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

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: September 27, 2017

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. When identification fails in one of these models, it becomes under-identified and the identification status of individual parameters is not generally straightforward to characterize. We provide a systematic reparameterization procedure that leads to a reparameterized model with straightforward identification status. Using this reparameterization, we determine the asymptotic behavior of 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. Using the asymptotic results, we propose 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 illustrates these results in three examples: a sample selection model, a triangular threshold crossing model and a collective model for household expenditures.

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 the identified set for the entire parameter vector is characterized by a lower-dimensional manifold inside of the vector's parameter space. Such a non-identification scenario is sometimes referred to as "under-identification" or "partial 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 provide a systematic reparameterization procedure that nonlinearly transforms a model's parameters into a new set of parameters that have straightforward identification status when identification fails. Second, using this reparameterization, we derive the limit theory for a class of standard extremum estimators (e.g., generalized method of moments, minimum distance and some forms of maximum likelihood) 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

---

[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.

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-dimensional subvectors, that are uniformly valid regardless of identification strength.

A substantial portion of the recent econometrics literature has been devoted to studying estimation in the presence of weak identification and developing inference tools that are robust to the identification strength of the parameters in an underlying economic or statistical model. Earlier papers in this line of research focus upon the linear instrumental variables (IV) 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, researchers have similarly characterized the behavior of standard estimators and inference procedures under various forms of weak identification (e.g., Stock and Wright, 2000) and developed robust inference procedures (e.g., Kleibergen, 2005). Most papers in this literature, such as Stock and Wright (2000) and Kleibergen (2005), focus 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) focuses 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, 2013a, 2014a) 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 restrict 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, nonlinear regression models, nonlinear IV models, certain dynamic stochastic general equilibrium (DSGE) models and structural vector autoregressions (VARs) identified by instruments or conditional heteroskedasticity. Indeed, this feature of a singular Jacobian without zero columns at points of identification failure is typical of many nonlinear 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 (2016b) in the context of minimum distance (MD) estimation and Andrews and Guggenberger (2014) and Andrews and Mikusheva (2016a) in the context of moment condition models. Andrews and Mikusheva (2016b) provide methods to directly perform uniformly valid subvector inference while Andrews and Guggenberger (2014) and Andrews and Mikusheva (2016a) do not.[2] Unlike these papers, but like Andrews and Cheng (2012a, 2013a, 2014a), we focus 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 (2016b). Also unlike these papers, but like Andrews and Cheng (2012a, 2013a, 2014a), 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 we study here is that the asymptotic size of standard Wald tests for the full parameter vector (and certain subvectors) *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 study various types of under-identification in different contexts. For example, Sargan (1983) studies regression models that are nonlinear in parameters and first-order locally under-identified. Phillips (1989) studies under-identified simultaneous equations models and spurious time series regressions. In a rather different context, Lee and Chesher (1986) also make use of a reparameterization for a type of identification problem. Arellano et al. (2012) proposes a way to test for under-identification in a generalized method of moments (GMM) context. Qu and Tkachenko (2012) study under-identification in the context of DSGE models. Escanciano and Zhu (2013) study under-identification in a class of semi-parametric models.[3] Dovonon and Renault (2013) uncover 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

---

[2] Andrews and Mikusheva (2016a) 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) or Andrews and Mikusheva (2016a) 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.

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 providing a general recipe for reparameterizing the extremum estimation problem so that, after reparameterization, it falls under the framework of Andrews and Cheng (2012a) (AC12 hereafter). More specifically, the reparameterization procedure involves solving a system of differential equations so that a set of the derivatives of the function that generates the reparameterization are in the null space of the Jacobian of the original model restrictions. This reparameterization generates a Jacobian of transformed model restrictions with zero columns at points of identification failure. This systematic approach to nonlinear reparameterization generalizes some antecedents in linear models for which the reparameterizations amount to linear rotations (e.g., Phillips, 1989). We show that the reparameterized extremum objective function satisfies a crucial assumption of AC12: at points of identification failure, it does not depend upon the unidentified parameters.[4] This allows us to use the results of AC12 to find the limit theory for the reparameterized parameter estimates.

We subsequently derive the limit theory for the original parameter estimates of economic interest using the fact that they are equal to a bijective function of the reparameterized parameter estimates. To obtain a full asymptotic characterization of the original parameter estimator, we rotate its subvectors in different directions of the parameter space. The subvector estimates converge 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) use similar rotations to derive limit theory for estimators under identification failure; Antoine and Renault (2009, 2012) use similar rotations to derive limit theory for estimators under "nearly-weak" identification;[5] Andrews and Cheng (2014a) (AC14 hereafter) use similar rotations to find the asymptotic distributions of Wald statistics under weak and nearly-strong identification; and recently Phillips (2016) uses similar rotations to find limit theory for regression estimators in the presence of near-multicollinearity in regressors. However, unlike their predecessors used for specific linear models, our nonlinear reparameterizations are not generally equivalent to the rotations we use to derive asymptotic theory.

We also derive the asymptotic distributions of standard Wald statistics for general (possibly

---

[4]This corresponds to Assumption A of AC12.

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

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. To overcome this issue, 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 (2017), 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. We then briefly analyze the power performance of one of our proposed robust Wald tests in a triangular threshold crossing model with a dummy endogenous variable. Finally, we apply the testing method in an empirical example that analyzes the effects of educational attainment on criminal activity.

The paper is organized as follows. In the next section, we introduce the general class of models subject to under-identification that we study and detail four examples of models in this class. Section 3 introduces a new method of systematic nonlinear reparameterization that leads to straightforward identification status under identification failure. This section includes a step-by-step algorithm for obtaining the reparameterization. Section 4 provides the limit theory for a general class of extremum estimators of the original model parameters under a comprehensive class of identification strengths. The nonstandard limit distributions derived here provide accurate approximations to the finite sample distributions of the parameter estimators, uncovered via Monte Carlo simulation. Section 5 similarly provides the analogous limit theory for standard Wald statistics. We describe how to perform uniformly robust inference in Section 6. Section 7 contains further details for a triangular threshold crossing model, including Monte Carlo simulations demonstrating how well the nonstandard limit distributions derived in Sections 4–5 approximate their finite-sample counterparts and an analysis of the power properties of a robust Wald test. Section 8 contains the empirical application. Proofs of the main results of the paper are provided in Appendix A, verification of assumptions for the threshold crossing model are contained in Appendix B, while figures are collected at the end of the document.

Notationally, we respectively let b_j, b^j and d_b denote the jth entry, the jth subvector and the dimension of a generic parameter vector b. All vectors in the paper are column vectors. However, to simplify notation, we occasionally abuse it by writing (c, d) instead of $(c', d')'$, $(c', d)'$ or $(c, d')'$ for vectors c and d and for a function $f(a)$ with $a = (c, d)$, we sometimes write $f(c, d)$ rather than $f(a)$.

2 General Class of Models

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

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

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

$$0 = \boldsymbol{g}^*(\boldsymbol{\theta}) = \xi^* - \mathfrak{g}(\boldsymbol{\theta}) \in \mathbb{R}^{d_g} \tag{2.2}$$

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

Often, econometric models imply a decomposition of $\boldsymbol{\theta}$: $\boldsymbol{\theta} = (\beta, \boldsymbol{\mu})$, where the parameter β determines the "identification status" of $\boldsymbol{\mu}$. That is, when $\beta \neq \bar{\beta}$ for some $\bar{\beta}$, $\boldsymbol{\mu}$ is identified; when $\beta = \bar{\beta}$, $\boldsymbol{\mu}$ is under-identified; and when β is "close" to $\bar{\beta}$ relative to sampling variability, $\boldsymbol{\mu}$ is local-to-under-identified. For convenience and without loss of generality, we use the normalization $\bar{\beta} = 0$. In this paper, we characterize identification of $\boldsymbol{\mu}$ via the Jacobian of the model restrictions:

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

The Jacobian $J^*(\boldsymbol{\theta})$ will have deficient rank across the subset of the parameter space for $\boldsymbol{\theta}$ for which $\beta = 0$ but full rank over the remainder of the parameter space.[6] Roughly speaking, we are considering models that become first-order under-identified in certain regions of the parameter space. Our main focus is on models for which the column rank of $J^*(\boldsymbol{\theta})$ lies strictly between 0 and $d_{\boldsymbol{\mu}}$ when $\beta = 0$ and this rank-deficiency is not the consequence of zero columns in $J^*(\boldsymbol{\theta})$; see Remark 3.1 below for a related discussion in terms of the information matrix. Although our results cover cases for which $J^*(\boldsymbol{\theta})$ has columns of zeros when $\beta = 0$, these cases are not of primary interest for this paper since they are nested in the framework of AC12.

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

---

[6]Assumption ID below is related to the former, and Assumption B3(iii) in AC12, which we assume later, implies the latter.

Remark 2.1. *For some models, we can further decompose $\boldsymbol{\theta}$ into $\boldsymbol{\theta} = (\beta, \boldsymbol{\mu}) = (\beta, \boldsymbol{\zeta}, \boldsymbol{\pi})$, where only the identification status of the subvector parameter $\boldsymbol{\pi}$ of $\boldsymbol{\mu}$ is affected by the value of β. More formally, when $\beta = 0$, $\mathrm{rank}\left(\partial \boldsymbol{g}^*(\boldsymbol{\theta})/\partial \boldsymbol{\pi}'\right) < d_{\boldsymbol{\pi}}$ for all $\boldsymbol{\theta} = (0, \boldsymbol{\zeta}, \boldsymbol{\pi}) \in \boldsymbol{\Theta}$ and $\boldsymbol{\gamma}^* \in \boldsymbol{\Gamma}$, where $\boldsymbol{\Theta}$ and $\boldsymbol{\Gamma}$ denote the parameter spaces of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. Modulo the reordering of the elements of $\boldsymbol{\mu}$, we can formalize the decomposition $\boldsymbol{\mu} = (\boldsymbol{\zeta}, \boldsymbol{\pi})$ as follows: $\boldsymbol{\pi}$ is the smallest subvector of $\boldsymbol{\mu}$ such that*

$$d_{\boldsymbol{\pi}} - rank\left(\partial \boldsymbol{g}^*(\boldsymbol{\theta})/\partial \boldsymbol{\pi}'\right) = d_{\boldsymbol{\mu}} - rank(J^*(\boldsymbol{\theta}))$$

when $\beta = 0$. That is, the rank deficiency of the Jacobian with respect to the subvector $\boldsymbol{\pi}$ is equal to the rank deficiency of the Jacobian with respect to the vector $\boldsymbol{\mu}$ when $\beta = 0$. This feature holds for Examples 2.1–2.3 below, and will be illustrated as a special case throughout the paper.

Example 2.1 (Sample selection models using the control function approach)**.**

$$Y_i = X_i'\boldsymbol{\pi}^1 + \varepsilon_i, \qquad D_i = \mathbf{1}[\boldsymbol{\zeta} + Z_{1i}'\beta \geq \nu_i],$$
$$(\varepsilon_i, \nu_i)' \sim F_{\varepsilon\nu}(\varepsilon, \nu; \boldsymbol{\pi}^2),$$

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 may include (components of) X_i. We observe $W_i = (D_iY_i, D_i, X_i, Z_i)$ and $F_{\varepsilon\nu}(\cdot, \cdot; \boldsymbol{\pi}^2)$ is a parametric distribution of the unobservable variables (ε, ν) parameterized by the scalar $\boldsymbol{\pi}^2$. The mean and variance of each unobservable is normalized to be zero and one, respectively. Constructing a moment condition based on the control function approach ([Heckman, 1979](#)), we have, when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$,

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

where $\boldsymbol{\theta} = (\beta, \boldsymbol{\zeta}, \boldsymbol{\pi}^1, \boldsymbol{\pi}^2)$ and the moment function is

$$\varphi(w, \boldsymbol{\theta}) = \begin{bmatrix} d \begin{bmatrix} x \\ \tilde{q}(\boldsymbol{\zeta} + z_1'\beta; \boldsymbol{\pi}^2) \end{bmatrix} \left[y - x'\boldsymbol{\pi}^1 - \tilde{q}(\boldsymbol{\zeta} + z_1'\beta; \boldsymbol{\pi}^2) \right] \\ \tilde{q}(\boldsymbol{\zeta} + z_1'\beta; \boldsymbol{\pi}^2)F_\nu^{-1}(-\boldsymbol{\zeta} - z_1'\beta)\left[d - F_\nu(\boldsymbol{\zeta} + z_1'\beta)\right]z \end{bmatrix}, \tag{2.4}$$

with $w = (dy, d, x, z)$ and $\tilde{q}(\cdot; \pi_2)$ being a known function. When $F_{\varepsilon\nu}(\varepsilon, \nu; \boldsymbol{\pi}^2)$ is a bivariate standard normal distribution with correlation coefficient $\boldsymbol{\pi}^2$, we have $F_\nu(\cdot) = \Phi(\cdot)$ and $\tilde{q}(\cdot; \boldsymbol{\pi}^2) = \boldsymbol{\pi}^2 q(\cdot)$ where $q(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the inverse Mill's ratio based on the standard normal density and distribution functions $\phi(\cdot)$ and $\Phi(\cdot)$.

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

$$
\begin{aligned}
Y_{1i} &= X_i'\boldsymbol{\pi}^1 + \varepsilon_{1i}, \\
Y_{0i} &= X_i'\boldsymbol{\pi}^2 + \varepsilon_{0i}, \qquad D_i = \mathbf{1}[\boldsymbol{\zeta} + Z_{1i}'\beta \geq \nu_i], \\
Y_i &= D_i Y_{1i} + (1 - D_i)Y_{0i}, \\
(\varepsilon_{1i}, \varepsilon_{0i}, \nu_i)' &\sim F_{\varepsilon_1,\varepsilon_0,\nu}(\varepsilon_1, \varepsilon_0, \nu; \boldsymbol{\pi}^3),
\end{aligned}
$$

where $F_{\varepsilon_1,\varepsilon_0,\nu}(\cdot, \cdot, \cdot; \boldsymbol{\pi}^3)$ is a parametric distribution of the unobserved variables $(\varepsilon_1, \varepsilon_0, \nu)$ parameterized by vector $\boldsymbol{\pi}^3$. We observe $W_i = (Y_i, D_i, X_i, Z_i)$. 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.

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

$$
\begin{aligned}
Y_i &= \mathbf{1}[\boldsymbol{\pi}_1 + \tilde{\boldsymbol{\pi}}_2 D_i - \varepsilon_i \geq 0] \\
D_i &= \mathbf{1}[\boldsymbol{\zeta} + \beta Z_i - \nu_i \geq 0]
\end{aligned} \quad, \quad (\varepsilon_i, \nu_i)' \sim F_{\varepsilon\nu}(\varepsilon_i, v_i; \boldsymbol{\pi}_3).
$$

where $Z_i \in \{0, 1\}$. We observe $W_i = (Y_i, D_i, Z_i)$. 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 for simplicity only. With $F_{\varepsilon\nu}(\varepsilon, \nu; \boldsymbol{\pi}_3) = \Phi(\varepsilon, \nu; \boldsymbol{\pi}_3)$, a bivariate standard normal distribution with correlation coefficient $\boldsymbol{\pi}_3$, the model becomes the usual bivariate probit model. A more general model with $F_{\varepsilon\nu}(\varepsilon, \nu; \boldsymbol{\pi}_3) = C(F_\varepsilon(\varepsilon), F_\nu(\nu); \boldsymbol{\pi}_3)$, for $C(\cdot, \cdot; \boldsymbol{\pi}_3)$ in a class of single parameter copulas, is considered in Han and Vytlacil (2017), whose generality we follow here. Let $\boldsymbol{\pi}_2 \equiv \boldsymbol{\pi}_1 + \tilde{\boldsymbol{\pi}}_2$ and, for simplicity, let F_ν and F_ε be uniform distributions.[7] The results of Han and Vytlacil (2017) provide that when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, $\xi^* - \mathbf{g}(\boldsymbol{\theta}) = 0$, where $\xi = (p_{11,0}, p_{11,1}, p_{10,0}, p_{10,1}, p_{01,0}, p_{01,1})'$ with $p_{yd,z} \equiv \Pr_{\boldsymbol{\gamma}}[Y = y, D = d | Z = z]$ and

$$
\mathbf{g}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{p}_{11,0}(\boldsymbol{\theta}) \\ \boldsymbol{p}_{11,1}(\boldsymbol{\theta}) \\ \boldsymbol{p}_{10,0}(\boldsymbol{\theta}) \\ \boldsymbol{p}_{10,1}(\boldsymbol{\theta}) \\ \boldsymbol{p}_{01,0}(\boldsymbol{\theta}) \\ \boldsymbol{p}_{01,1}(\boldsymbol{\theta}) \end{bmatrix} \equiv \begin{bmatrix} C(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \\ C(\boldsymbol{\pi}_2, \boldsymbol{\zeta} + \beta; \boldsymbol{\pi}_3) \\ \boldsymbol{\pi}_1 - C(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \\ \boldsymbol{\pi}_1 - C(\boldsymbol{\pi}_1, \boldsymbol{\zeta} + \beta; \boldsymbol{\pi}_3) \\ \boldsymbol{\zeta} - C(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \\ \boldsymbol{\zeta} + \beta - C(\boldsymbol{\pi}_2, \boldsymbol{\zeta} + \beta; \boldsymbol{\pi}_3) \end{bmatrix}. \tag{2.5}
$$

---

[7]This normalization is not necessary and is only introduced here for simplicity; see Han and Vytlacil (2017) for the formulation of the identification problem without it.

For later use, we also define the (redundant) probabilities:

$$\boldsymbol{p}_{00,0}(\boldsymbol{\theta}) \equiv 1 - \boldsymbol{p}_{11,0}(\boldsymbol{\theta}) - \boldsymbol{p}_{10,0}(\boldsymbol{\theta}) - \boldsymbol{p}_{01,0}(\boldsymbol{\theta}), \tag{2.6}$$
$$\boldsymbol{p}_{00,1}(\boldsymbol{\theta}) \equiv 1 - \boldsymbol{p}_{11,1}(\boldsymbol{\theta}) - \boldsymbol{p}_{10,1}(\boldsymbol{\theta}) - \boldsymbol{p}_{01,1}(\boldsymbol{\theta}).$$

Example 2.4 (Engel curve models for household share). *Tommasi and Wolf (2016) discuss Engel curve estimation for the private assignable good in the Dunbar et al. (2013) collective model for household expenditure shares when using the PIGLOG utility function. See equation (5) of Tommasi and Wolf (2016) for these Engel curves. These authors estimate the model parameters by a particular nonlinear least squares criterion. We instead consider the general GMM estimation problem in this context for which $0 = \boldsymbol{g}^*(\boldsymbol{\theta}) = E_{\boldsymbol{\gamma}^*}\varphi(W_i, \boldsymbol{\theta})$ when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, where $\boldsymbol{\theta} = (\beta, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3)$ and the moment function is*

$$\varphi(w, \boldsymbol{\theta}) = A(y_h) \left[\begin{pmatrix} w_{1,h} \\ w_{2,h} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\pi}_1(\boldsymbol{\pi}_2 + \boldsymbol{\pi}_3 + \beta\log(\boldsymbol{\pi}_1 y_h)) \\ (1 - \boldsymbol{\pi}_1)(\boldsymbol{\pi}_2 + \beta\log((1 - \boldsymbol{\pi}_1)y_h)) \end{pmatrix} \right], \tag{2.7}$$

where $A(\cdot)$ is some $(d_{\boldsymbol{g}} \times 2)$-dimensional function. For example,

$$A(y_h) = \begin{bmatrix} 1 & 0 \\ y_h & 0 \\ 0 & 1 \\ 0 & y_h \end{bmatrix}.$$

There are many other examples of models that fit our framework including but not limited to nonlinear IV models, nonlinear regression models, certain DSGE models and structural VARs identified by conditional heteroskedasticity or instruments.

Examples 2.1 and 2.2 are contained in a class of moment condition models that uses a control function approach to account for endogeneity. This class of models fits our framework so that when $\beta = 0$, the control function loses its exogenous variability and the model presents multicollinearity in the Jacobian matrix. In Example 2.1, with $q(\cdot)$ being the inverse Mill's ratio, the Jacobian matrix (2.3) satisfies

$$J^*(\boldsymbol{\theta}) = E_{\boldsymbol{\gamma}^*} \begin{bmatrix} -\boldsymbol{\pi}^2 D_i X_i q_i' & -D_i X_i X_i' & -D_i q_i X_i \\ D_i Y_i q_i' - D_i X_i' \boldsymbol{\pi}^1 q_i' - 2\boldsymbol{\pi}^2 D_i q_i q_i' & -D_i q_i X_i' & -D_i q_i^2 \\ L_i(\beta, \boldsymbol{\zeta}) Z_i & 0_{l \times k} & 0_{l \times 1} \end{bmatrix},$$

where $q_i \equiv q(\boldsymbol{\zeta} + Z_{1i}'\beta)$, $q_i' \equiv dq(x)/dx|_{x=\boldsymbol{\zeta}+Z_{1i}'\beta}$,

$$L_i(\beta, \boldsymbol{\zeta}) \equiv \frac{\{q_i'(D_i - \Phi_i) - q_i\phi_i\}(1 - \Phi_i) + q_i\phi_i(D_i - \Phi_i)}{(1 - \Phi_i)^2}, \tag{2.8}$$

$\Phi_i \equiv \Phi(\boldsymbol{\zeta} + Z_{1i}'\beta)$ and $\phi_i \equiv \phi(\boldsymbol{\zeta} + Z_{1i}'\beta)$. Note that $d_{\boldsymbol{\zeta}} < rank(J^*(\boldsymbol{\theta})) < d_{\boldsymbol{\mu}}$ when $\beta = 0$, since q_i becomes a constant and $X_i = (1, X_{1i}')'$.

In general, a rank-deficient Jacobian with non-zero columns when $\beta = 0$ poses several challenges rendering existing asymptotic theory in the literature that considers a Jacobian with zero columns when identification fails inapplicable here: (i) since none of the columns of $J^*(\boldsymbol{\theta})$ are equal to zero, it is often unclear which components of the $\boldsymbol{\pi}$ parameter are (un-)identified; (ii) key assumptions in the literature, such as Assumption A in AC12, do not hold; (iii) typically, $\boldsymbol{g}^*(\boldsymbol{\theta})$ or $J^*(\boldsymbol{\theta})$ 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 Systematic Reparameterization

In this section, we define the criterion functions used for estimation and the sample model restriction functions that enter them and formally impose assumptions on these two objects. We then introduce a systematic method for reparameterizing general under-identified models. After reparameterization, the identification status of the model parameters becomes straightforward with individual parameters being either well identified or completely unidentified when identification fails. We later use this reparameterization procedure as a step toward obtaining limit theory for estimators and tests of the original parameters of interest under a comprehensive class of identification strengths. However, this reparameterization procedure carries some interest in its own right because it (i) characterizes the submanifold of the original parameter space that is (un)identified and (ii) has the potential for application to finding the limit theory for general globally under-identified models (in contrast to those that lose identification at points in the parameter space for which $\beta = 0$).

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

$$\hat{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta} \text{ and } \boldsymbol{Q}_n(\hat{\boldsymbol{\theta}}_n) = \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \boldsymbol{Q}_n(\boldsymbol{\theta}) + o(n^{-1}).$$

In the following assumptions we presume that $\boldsymbol{Q}_n(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$ only through the sample counterpart $\bar{\boldsymbol{g}}_n(\boldsymbol{\theta})$ of $\boldsymbol{g}^*(\boldsymbol{\theta})$. In the case of MD and some particular maximum likelihood (ML) models, $\bar{\boldsymbol{g}}_n(\boldsymbol{\theta}) = \hat{\xi}_n - \mathbf{g}(\boldsymbol{\theta})$, where $\hat{\xi}_n$ is a sample analog of ξ^*, in analogy to (2.2). For GMM,

$\bar{g}_n(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \varphi(W_i, \boldsymbol{\theta})$.

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

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

for some random function $\Psi_n(\cdot)$ *that is differentiable.*

Assumption CF is naturally satisfied when we construct GMM/MD or ML criterion functions, given (2.1) or (2.2). Note that models that generate minimum distance structures and the types of likelihoods that fall under our framework typically involve $\boldsymbol{g}^*(\boldsymbol{\theta}) = \xi^* - \mathbf{g}(\boldsymbol{\theta})$ by (2.2). For a GMM/MD criterion function, $\Psi_n(\bar{g}_n(\boldsymbol{\theta})) = \|W_n \bar{g}_n(\boldsymbol{\theta})\|^2$ where W_n is a (possibly random) weight matrix.[8] For a ML criterion function, $\Psi_n(\bar{g}_n(\boldsymbol{\theta})) = -\frac{1}{n} \sum_{i=1}^{n} \ln f^\dagger \left(W_i; \hat{\xi}_n - \bar{g}_n(\boldsymbol{\theta}) \right)$ if the distribution of the data depends on $\boldsymbol{\theta}^*$ only through $\xi^* = \mathbf{g}(\boldsymbol{\theta}^*)$, which is a reduced-form parameter (Rothenberg, 1971). That is, there exists a function $f^\dagger(w; \cdot)$ such that

$$f(w; \boldsymbol{\theta}) = f^\dagger(w; \xi^* - \boldsymbol{g}^*(\boldsymbol{\theta})) = f^\dagger(w; \mathbf{g}(\boldsymbol{\theta})),$$

where $f(\cdot; \boldsymbol{\theta})$ is the density of W_i. Of course, our framework may also accommodate ML estimation performed via a GMM criterion function that uses the score equations as moment vectors. However, the usual equivalence between this GMM estimation approach (using the efficient weighting matrix) and direct maximization of the likelihood function no longer holds in the weak identification scenarios considered in this paper.

Assumption Reg1. $\bar{g}_n : \boldsymbol{\Theta} \to \mathbb{R}^{d_g}$ *is continuously differentiable in* $\boldsymbol{\theta}$.

Assumption ID. *When* $\beta = 0$, $rank(\partial \bar{g}_n(\boldsymbol{\theta})/\partial \boldsymbol{\mu}') \equiv r < d_{\boldsymbol{\mu}}$ *for all* $\boldsymbol{\theta} = (0, \boldsymbol{\mu}) \in \boldsymbol{\Theta}$.

To simplify the asymptotic theory derived in Section 4, we impose the following assumption that ensures the reparameterization function $h(\cdot)$ in Procedure 3.1 below is nonrandom and does not depend on the true DGP.

Assumption Jac. *When* $\beta = 0$, *the null space of* $J^*(\boldsymbol{\theta})$ *is equal to the null space of* $\partial \bar{g}_n(\boldsymbol{\theta})/\partial \boldsymbol{\mu}'$ *for all* $n \geq 1$ *and does not depend upon* ϕ^*.

This assumption guarantees that the reparameterization we later obtain is deterministic and does not depend upon the true DGP. Example 2.1–2.4 satisfy this assumption. However, the asymptotic theory derived in Section 4 can be extended to some cases for which our reparameterization is random and/or DGP-dependent, but we have not found an application for which such an extension would be useful.

---

[8]Note that Assumption CF does not cover GMM with a continuously updating weight matrix $W_n(\boldsymbol{\theta})$.

Remark 3.1. *Given the existence of $f^\dagger(w; \cdot)$ in the ML framework, the setting of this paper can be characterized in terms of the information matrix. Let $\mathcal{I}(\boldsymbol{\theta})$ be the $d_{\boldsymbol{\theta}} \times d_{\boldsymbol{\theta}}$ information matrix*

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

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

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

and $\mathcal{I}^\dagger(\mathbf{g}) \equiv E\left(\partial \log f^\dagger/\partial \mathbf{g}\right)\left(\partial \log f^\dagger/\partial \mathbf{g}'\right)$ has full rank.[9]

We now propose a systematic reparameterization as a key step toward deriving the limit theory under various strengths of identification. Let d_π denote the rank reduction in the sample Jacobian $\partial \bar{g}_n(\boldsymbol{\theta})/\partial \boldsymbol{\mu}$ under identification failure, i.e., $d_\pi \equiv d_{\boldsymbol{\mu}} - r$ (this will later denote the dimension of a new parameter π). Let the parameter space for $\boldsymbol{\mu}$ be denoted as

$$\mathcal{M} = \{\boldsymbol{\mu} \in \mathbb{R}^{d_{\boldsymbol{\mu}}} : \boldsymbol{\theta} = (\beta, \boldsymbol{\mu}) \text{ for some } \boldsymbol{\theta} \in \boldsymbol{\Theta}\}.$$

The reparameterization procedure in its most general form proceeds in two steps:

Procedure 3.1. *For a given $\bar{g}_n(\boldsymbol{\theta})$ that satisfies Assumptions Reg1 and ID, let $\theta = (\beta, \mu)$ denote a new vector of parameters for which $d_\mu = d_{\boldsymbol{\mu}}$. Find a reparameterization function $h(\cdot)$ as follows:*

1. *Find a deterministic full rank $d_{\boldsymbol{\mu}} \times d_{\boldsymbol{\mu}}$ matrix M that performs elementary column operations*[10] *such that when $\beta = 0$,*

$$\frac{\partial \bar{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}'} M(\boldsymbol{\mu}) = \left[\boldsymbol{G}_n(\boldsymbol{\mu}) : 0_{d_{\boldsymbol{g}} \times d_\pi}\right] \tag{3.1}$$

   for all $\boldsymbol{\mu} \in \mathcal{M}$, where $\boldsymbol{G}_n(\boldsymbol{\mu})$ is some $d_{\boldsymbol{g}} \times r$ matrix.[11]

2. *Find a differentiable one-to-one function $h : \mathcal{M} \to \mathcal{M}$ such that*

$$\frac{\partial h(\mu)}{\partial \mu'} = M(h(\mu))$$

---

[9]This is because $\xi = \mathbf{g}(\boldsymbol{\theta})$ is a reduced-form parameter that is always (strongly) identified.

[10]There are three types of elementary column operations: switching two columns, multiplying a column with a non-zero constant, and replacing a column with the sum of that column and a multiple of another column.

[11]The existence of such a matrix M is guaranteed by Assumption Jac.

for all $\mu \in \mathcal{M}$, where

$$\mathcal{M} \equiv \{\mu \in \mathbb{R}^{d_\mu} : \boldsymbol{\theta} = (\beta, h(\mu)) \text{ for some } \boldsymbol{\theta} \in \boldsymbol{\Theta}\}.$$

Proposition 3.1 below provides sufficient conditions for the existence of a $h(\cdot)$ function resulting from Procedure 3.1. We also note that the singular value decomposotion can be used to compute the matrix $M(\boldsymbol{\mu})$ with conventional software since the right singular vectors of $\partial \bar{\boldsymbol{g}}_n(\boldsymbol{\theta})/\partial \boldsymbol{\mu}'$ that correspond to its zero singular values span its null space and its left singular vectors that correspond to its non-zero singular values span its column space.[12] With the reparameterization function $h(\cdot)$, we transform $\boldsymbol{\mu}$ to μ such that $\boldsymbol{\mu} = h(\mu)$. That is, we have the reparameterization as the following one-to-one map:

$$\theta \equiv (\beta, \mu) \mapsto \boldsymbol{\theta} \equiv (\beta, \boldsymbol{\mu}), \tag{3.2}$$

where $(\beta, \boldsymbol{\mu}) = (\beta, h(\mu))$. Let π denote the subvector composed of the final d_π entries of the new parameter μ so that we may write $\mu = (\zeta, \pi)$. We illustrate this reparameterization approach in the following continuation of Example 2.1. The approach is further illustrated in Examples 2.3–2.4 below.

Examples 2.1 and 2.2, continued. *Since Examples 2.1 and 2.2 are similar in the aspects we focus on, we only analyze Example 2.1 in further detail. In this example, we are considering a GMM estimator so that $\bar{\boldsymbol{g}}_n(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^n \varphi(W_i, \boldsymbol{\theta})$, where the moment function $\varphi(w, \boldsymbol{\theta})$ is given by (2.4). In the case for which $F_{\varepsilon\nu}(\varepsilon, \nu; \boldsymbol{\pi}^2)$ is a bivariate standard normal distribution, the sample Jacobian for this model with respect to $\boldsymbol{\mu}$ is*

$$\frac{\partial \bar{\boldsymbol{g}}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}'} = -\frac{1}{n} \sum_{i=1}^n \begin{bmatrix} \boldsymbol{\pi}^2 D_i X_i q'(\boldsymbol{\zeta}) & D_i X_i X_i' & D_i q(\boldsymbol{\zeta}) X_i \\ D_i X_i' \boldsymbol{\pi}^1 q'(\boldsymbol{\zeta}) + (2\boldsymbol{\pi}^2 q(\boldsymbol{\zeta}) - Y_i) D_i q'(\boldsymbol{\zeta}) & D_i q(\boldsymbol{\zeta}) X_i' & D_i q(\boldsymbol{\zeta})^2 \\ -L_i(0, \boldsymbol{\zeta}) Z_i & 0_{l \times k} & 0_{l \times 1} \end{bmatrix}$$

when $\beta = 0$, where $L_i(\beta, \boldsymbol{\zeta})$ is defined in (2.8). Note that $r = d_{\boldsymbol{\mu}} - 1$ since the final column is a scalar multiple of the $(l + 1)^{th}$ so that $d_\pi = 1$. For Step 1 of Procedure 3.1, we set the final column of $M(\boldsymbol{\mu})$ equal to $(0, -q(\boldsymbol{\zeta}), 0_{1 \times (k-1)}, 1)'$. For Step 2, we find the general solution in $h(\cdot)$ to the following system of ODEs:

$$\frac{\partial h(\mu)}{\partial \pi} = (0, -q(h_1(\mu)), 0_{1 \times (k-1)}, 1)'.$$

---

[12]We thank Áureo de Paula for pointing this out.

13

This yields

$$h(\mu) = (c^1(\zeta), -q(c^1(\zeta))\pi + c^2(\zeta), c^3(\zeta)', \pi + c^4(\zeta))',$$

where $c^1(\zeta)$, $c^2(\zeta)$ and $c^4(\zeta)$ are arbitrary one-dimensional constants of integration that may depend on ζ and $c^3(\zeta)$ is an arbitrary $(k-1)$-dimensional constant of integration that may depend on ζ. Upon setting $c^1(\zeta) = \zeta_1$, $c^2(\zeta) = \zeta_2$, $c^3(\zeta) = (\zeta_3, \ldots, \zeta_{k+1})'$ and $c^4(\zeta) = 0$, we have

$$\frac{\partial h(\mu)}{\partial \mu'} = \begin{bmatrix} 1 & 0 & 0_{1\times(k-1)} & 0 \\ -q'(\zeta_1)\pi & 1 & 0_{1\times(k-1)} & -q(\zeta_1) \\ 0_{(k-1)\times 1} & 0_{(k-1)\times 1} & I_{k-1} & 0_{(k-1)\times 1} \\ 0 & 0 & 0_{1\times(k-1)} & 1 \end{bmatrix}$$

being full rank. Thus, we have found a one-to-one reparameterization function $h(\cdot)$ such that $\boldsymbol{\mu} = (\boldsymbol{\zeta}, \boldsymbol{\pi}) = h(\mu) = (\zeta_1, \zeta_2 - q(\zeta_1)\pi, \zeta_3, \ldots, \zeta_{k+1}, \pi)$, or equivalently, $\zeta_1 = \boldsymbol{\zeta}$, $\zeta_2 = \boldsymbol{\pi}_1^1 + q(\boldsymbol{\zeta})\boldsymbol{\pi}^2$, $(\zeta_3, \ldots, \zeta_{k+1}) = (\boldsymbol{\pi}_2^1, \ldots, \boldsymbol{\pi}_k^1)$ and $\pi = \boldsymbol{\pi}^2$.

Define the sample model restriction and the criterion functions of the new parameter θ as

$$\bar{g}_n(\theta) \equiv \bar{\boldsymbol{g}}_n(\beta, h(\mu))$$

and

$$Q_n(\theta) \equiv \boldsymbol{Q}_n(\beta, h(\mu)).$$

The new Jacobian $\partial \bar{g}_n(\theta)/\partial \mu' = (\partial \bar{\boldsymbol{g}}_n(\boldsymbol{\theta})/\partial \boldsymbol{\mu}')(\partial h(\mu)/\partial \mu')$ has the same reduced rank $r < d_\mu = d_{\boldsymbol{\mu}}$ as the original Jacobian $\partial \bar{\boldsymbol{g}}_n(\boldsymbol{\theta})/\partial \boldsymbol{\mu}'$ since $\partial h(\mu)/\partial \mu = M(h(\mu))$ has full rank. But now, by the construction of the reparameterization function $h(\cdot)$ according to Procedure 3.1, the rank reduction arises purely from the final d_π columns of $\partial \bar{g}_n(\theta)/\partial \mu$ being equal to zero. Using this result, the reparameterized criterion function $Q_n(\theta)$ satisfies a property that is instrumental to deriving the limit theory detailed below.

Theorem 3.1. *Under Assumptions CF, Reg1 and ID, $Q_n(\theta)$ does not depend upon π when $\beta = 0$ for all $\theta = (0, \zeta, \pi) \in \Theta$.*

In conjunction with other assumptions, the result of this theorem allows us to apply the asymptotic results in Theorems 3.1 and 3.2 of AC12 to the *reparameterized criterion function* $Q_n(\theta)$, the *new parameter* θ and estimator $\hat{\theta}_n$, defined by

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

where Θ is the optimization parameter space in the reparameterized estimation problem and is

defined in terms of the original optimization parameter space $\boldsymbol{\Theta}$ as follows:

$$\Theta \equiv \{(\beta, \mu) \in \mathbb{R}^{d_\theta} : (\beta, h(\mu)) \in \boldsymbol{\Theta}\}.$$

We now provide an algorithm for practical implementation of Procedure 3.1.

Algorithm 3.1. *For a given $\bar{\boldsymbol{g}}_n(\boldsymbol{\theta})$ that satisfies Assumptions Reg1 and ID, let $\theta = (\beta, \mu) = (\beta, \zeta, \pi)$ denote a new vector of parameters for which $d_\mu = d_{\boldsymbol{\mu}}$. Find a reparameterization function $h(\cdot)$ as follows:*

1. *Find a deterministic non-zero $d_{\boldsymbol{\mu}} \times 1$ vector $m^{(1)}$ such that when $\beta = 0$,*

$$\frac{\partial \bar{\boldsymbol{g}}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}'} m^{(1)}(\boldsymbol{\mu}) = 0_{d_{\boldsymbol{g}} \times 1} \tag{3.3}$$

    for all $\boldsymbol{\mu} \in \mathcal{M}$.

2. *Let $\mu^{(1)} = (\zeta^{(1)}, \pi^{(1)})$ denote a new $d_{\boldsymbol{\mu}} \times 1$ vector of parameters, where $\pi^{(1)}$ is a $d_\pi \times 1$ subvector. Find the general solution in $h^{(1)} : \mathcal{M}^{(1)} \to \mathcal{M}$ to the following system of first order ordinary differential equations (ODEs):*

$$\frac{\partial h^{(1)}(\mu^{(1)})}{\partial \pi_1^{(1)}} = m^{(1)}(h^{(1)}(\mu^{(1)})) \tag{3.4}$$

    for all $\mu^{(1)} \in \mathcal{M}^{(1)} \equiv \{\mu^{(1)} \in \mathbb{R}^{d_{\boldsymbol{\mu}}} : \boldsymbol{\theta} = (\beta, h^{(1)}(\mu^{(1)})) \text{ for some } \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$.

3. *From the general solution for $h^{(1)}$ in Step 2, find a particular solution for $h^{(1)}$ such that the matrix $\partial h^{(1)}(\mu^{(1)})/\partial \mu^{(1)\prime}$ has full rank for all $\mu^{(1)} \in \mathcal{M}^{(1)}$.[13]*

4. *If $d_\pi = 1$ (i.e., $\pi_1^{(1)} = \pi^{(1)}$), stop and set $h = h^{(1)}$ and $\mu = \mu^{(1)}$. Otherwise, set $\theta^{(1)} = (\beta, \mu^{(1)})$, $g_n^{(1)}(\theta^{(1)}) = \bar{\boldsymbol{g}}_n(\beta, h^{(1)}(\mu^{(1)}))$, $\Theta^{(1)} = \{(\beta, \mu^{(1)}) \in \mathbb{R}^{d_\theta} : (\beta, h^{(1)}(\mu^{(1)})) \in \boldsymbol{\Theta}\}$ and $i = 2$ (moving to the second iteration of the algorithm) and continue to the next step.*

5. *Find a non-zero $d_{\boldsymbol{\mu}} \times 1$ vector $m^{(i)}$ such that when $\beta = 0$,*

$$\frac{\partial g_n^{(i-1)}(\theta^{(i-1)})}{\partial \mu^{(i-1)\prime}} m^{(i)}(\mu^{(i-1)}) = 0_{d_{\boldsymbol{g}} \times 1} \tag{3.5}$$

    for all $\mu^{(i-1)} \in \mathcal{M}^{(i-1)}$.

---

[13]When evaluated at $\boldsymbol{\mu} = h^{(1)}(\mu^{(1)})$, the vector $m^{(1)}(\boldsymbol{\mu})$ is a column in the matrix $\partial h^{(1)}(\mu^{(1)})/\partial \mu^{(1)\prime}$, denoted as $M^{(1)}$ later. The analogous statement applies to $m^{(i)}$ in Steps 5–6. In the special case for which $d_\pi = 1$, $m^{(1)}(\boldsymbol{\mu})$ evaluated at $\boldsymbol{\mu} = h^{(1)}(\mu^{(1)})$ is equal to the final column of $\partial h^{(1)}(\mu^{(1)})/\partial \mu^{(1)\prime}$.

6. Let $\mu^{(i)} = (\zeta^{(i)}, \pi^{(i)})$ denote a new $d_{\boldsymbol{\mu}} \times 1$ vector of parameters, where $\pi^{(i)}$ is a $d_\pi \times 1$ subvector. Find the general solution in $h^{(i)} : \mathcal{M}^{(i)} \to \mathcal{M}^{(i-1)}$ to the following system of first order ODEs:

$$\frac{\partial h^{(i)}(\mu^{(i)})}{\partial \pi_i^{(i)}} = m^{(i)}(h^{(i)}(\mu^{(i)})), \tag{3.6}$$

for all $\mu^{(i)} \in \mathcal{M}^{(i)} \equiv \{\mu^{(i)} \in \mathbb{R}^{d_{\boldsymbol{\mu}}} : \theta^{(i-1)} = (\beta, h^{(i)}(\mu^{(i)}))$ for some $\theta^{(i-1)} \in \Theta^{(i-1)}\}$.

7. From the general solution for $h^{(i)}$ in Step 6, find a particular solution for $h^{(i)}$ such that for all $\mu^{(i)} \in \mathcal{M}^{(i)}$ (1) the matrix $\partial h^{(i)}(\mu^{(i)})/\partial \mu^{(i)\prime}$ has full rank and (2)

$$\frac{\partial h^{(i)}(\mu^{(i)})}{\partial (\pi_1^{(i)}, ..., \pi_{i-1}^{(i)})} = \left[\begin{array}{c} 0_{(d_{\boldsymbol{\mu}} - d_\pi) \times (i-1)} \\ C^{(i)}(\mu^{(i)}) \\ 0_{(d_\pi - i + 1) \times (i-1)} \end{array} \right],$$

where $C^{(i)}(\mu^{(i)})$ is an arbitrary $(i-1) \times (i-1)$ matrix.

8. If $i = d_\pi$, stop and set $h = h^{(1)} \circ \ldots \circ h^{(d_\pi)}$ and $\mu = \mu^{(d_\pi)}$. Otherwise, set $\theta^{(i)} = (\beta, \mu^{(i)})$, $g_n^{(i)}(\theta^{(i)}) = g_n^{(i-1)}(\beta, h^{(i)}(\mu^{(i)}))$, $\Theta^{(i)} = \{(\beta, \mu^{(i)}) \in \mathbb{R}^{d_\theta} : (\beta, h^{(i)}(\mu^{(i)})) \in \Theta^{(i-1)}\}$ and $i = i + 1$ and return to Step 5.

As is the case for Procedure 3.1, the function $h(\cdot)$ is a reparameterization function that maps the new parameter μ to the original parameter $\boldsymbol{\mu}$ in accordance with (3.2), i.e., $\boldsymbol{\mu} = h(\mu)$. We formally establish the connection between Algorithm 3.1 and Procedure 3.1.

Theorem 3.2. *Define* $\mathcal{M} = \mathcal{M}^{(d_\pi)}$, *where* $\mathcal{M}^{(d_\pi)}$ *is defined in Step 6 of Algorithm 3.1. The reparameterization function* $h : \mathcal{M} \to \boldsymbol{\mathcal{M}}$ *constructed according to Algorithm 3.1 constitutes a solution to Procedure 3.1.*

Remark 3.2. *Defining the matrix function* $M^{(i)}(h^{(i)}(\mu^{(i)})) = \partial h^{(i)}(\mu^{(i)})/\partial \mu^{(i)\prime}$ *for* $i = 1, \ldots, d_\pi$ *consistently with the notation used in Algorithm 3.1 so that each* $m^{(i)}(h^{(i)}(\mu^{(i)}))$ *is the* $(d_\zeta + i)^{th}$ *column of* $M^{(i)}(h^{(i)}(\mu^{(i)}))$, *we note that the matrix performing elementary operations in Procedure 3.1 can be expressed as*

$$M(h(\mu)) = M^{(1)}(h^{(1)} \circ \ldots \circ h^{(d_\pi)}(\mu)) \times \ldots \times M^{(d_\pi)}(h^{(d_\pi)}(\mu)).$$

We also note that in terms of the recursive parameter spaces of Algorithm 3.1, $\Theta = \Theta^{(d_\pi)}$.

When implementing Steps 3 and 7 of Algorithm 3.1, knowledge of the well-identified parameter $\boldsymbol{\zeta}$ in $\boldsymbol{\mu} = (\boldsymbol{\zeta}, \boldsymbol{\pi})$ is useful in making $\partial h^{(i)}(\mu^{(i)})/\partial \zeta^{(i)}$ relatively simple; see Remark 3.5

and the examples below. We note that the reparameterizations resulting from Procedure 3.1 or Algorithm 3.1 are not necessarily unique though such non-uniqueness poses no problems for our analysis. A sufficient condition for the existence of such a reparameterization is provided as follows.

Assumption Lip. $m^{(i)}(\cdot)$ *is Lipschitz continuous on compact* $\mathcal{M}^{(i-1)}$ *for every* $i = 1, ..., d_\pi$ *with* $\mathcal{M}^{(0)} \equiv \mathcal{M}$.

Proposition 3.1. *Under Assumptions Reg1, ID and Lip, there exists a reparametrization function* $h(\cdot)$ *on* $\mathcal{M}$ *that is an output of Algorithm 3.1 if Assumption Lip holds.*

Assumption Lip is related to restrictions on $\bar{g}(\boldsymbol{\theta})$. In practice, one can verify this assumption by simply calculating $m^{(i)}(\cdot)$ in Step 2 or 5 in Algorithm 3.1, as these steps are straightforward to implement.

Remark 3.3. *The nonlinear reparameterization approach we pursue here results in a new parameter with straightforward identification status when identification fails:* ζ *is well-identified and* π *is completely unidentified. When* β *is close to zero,* π *will be weakly identified while* (β, ζ) *remain strongly identified. Our analysis can be seen as a generalization of linear rotation-based reparameterization approaches that have been successfully used to transform linear models in the presence of identification failure so that the new parameters have the same straightforward identification status. See for example, Phillips (1989) in the context of linear IV models and Phillips (2016) in the context of the linear regression model with potential multicollinearity.*

Remark 3.4. *We note that our systematic reparameterization approach may also be useful in contexts for which a particular model is globally under-identified across its entire parameter space (not just in the region for which a parameter* β *is equal to zero). The reparameterization procedure may be useful for analyzing the identification properties of such models as well as determining the limiting behavior of parameter estimates and test statistics. For globally under-identified models with a constant (deficient) rank Jacobian, the subsequent results of sections 4–6 could be modified so that no parameter* β *appears in the analysis and the relevant limiting distributions would correspond to those derived under weak identification with the localization parameter b simply set equal to zero. For example, such an approach may be useful for under-identified DSGE models used in macroeconomics (see e.g., Komunjer and Ng, 2011 and Qu and Tkachenko, 2012). Further analysis of this approach is well beyond the scope of the present paper.*

Remark 3.5. *As can be seen from the continuation of Examples 2.1 and 2.3, when we know the component* $\boldsymbol{\zeta}$ *of* $\boldsymbol{\mu}$ *is well-identified for all values of* β, *we can form* $h(\cdot)$ *so that the first*

$d_{\boldsymbol{\zeta}}$ elements of $h(\mu)$ are equal to the first $d_{\boldsymbol{\zeta}}$ elements of the new well-identified parameter $\zeta = (\zeta^1, \zeta^2)$, viz., $\boldsymbol{\zeta} = (h_1(\mu), \ldots, h_{d_{\boldsymbol{\zeta}}}(\mu)) = \zeta^1$. In this special case, the reparameterization (3.2) can be written as a one-to-one map

$$\theta \equiv (\beta, \zeta, \pi) \mapsto \boldsymbol{\theta} \equiv (\beta, \boldsymbol{\zeta}, \boldsymbol{\pi}),$$

where $(\beta, \boldsymbol{\zeta}, \boldsymbol{\pi}) = (\beta, \zeta^1, h^2(\zeta^2, \pi))$ with $\mu = (\zeta^1, \zeta^2, \pi) = (\zeta, \pi)$ and $\boldsymbol{\zeta}$ is the new always well-identified parameter.

We close this section by illustrating the reparameterization algorithm with two other examples discussed earlier.

Examples 2.3, continued. *Given the specification of a single parameter copula $C(\cdot, \cdot; \boldsymbol{\pi}_3)$, this model can be estimated by minimizing the negative (conditional) likelihood function so that $\bar{g}_n(\boldsymbol{\theta}) = \hat{\xi}_n - \mathbf{g}(\boldsymbol{\theta})$, where $\hat{\xi}_n$ is equal to a vector of the empirical probabilities corresponding to the $p_{yd,z}$'s and $\mathbf{g}(\boldsymbol{\theta})$ is defined in (2.5).[14] The sample Jacobian for this model with respect to $\boldsymbol{\mu}$ is*

$$\frac{\partial \bar{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}'} = -\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}'} = - \begin{bmatrix} C_2(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & 0 & C_1(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & C_3(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \\ C_2(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & 0 & C_1(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & C_3(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \\ -C_2(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & 1 - C_1(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & 0 & -C_3(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \\ -C_2(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & 1 - C_1(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & 0 & -C_3(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \\ 1 - C_2(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & 0 & -C_1(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & -C_3(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \\ 1 - C_2(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & 0 & -C_1(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & -C_3(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \end{bmatrix}$$

when $\beta = 0$, where $C_1(\cdot, \cdot; \boldsymbol{\pi}_3)$, $C_2(\cdot, \cdot; \boldsymbol{\pi}_3)$ and $C_3(\cdot, \cdot; \boldsymbol{\pi}_3)$ denote the derivatives of $C(\cdot, \cdot; \boldsymbol{\pi}_3)$ with respect to the first argument, the second argument and $\boldsymbol{\pi}_3$. This matrix contains only three linearly independent row so that $r = d_{\boldsymbol{\mu}} - 1$. In the following analysis, since $d_\pi = 1$, we simplify notation by letting $h^{(1)} = h$, $m^{(1)} = m$ and $\mu^{(1)} = \mu = (\zeta, \pi)$. For Step 1 of Algorithm 3.1, we set $m(\boldsymbol{\mu}) = (0, C_3(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3)/(1 - C_1(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3)), -C_3(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3)/C_1(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3), 1)'$. For Step 2,

---

[14]Maximizing the conditional likelihood is equivalent to maximizing the full likelihood for this problem.

a set of general solutions to the system of ODEs

$$\frac{\partial h(\mu)}{\partial \pi} = \begin{pmatrix} 0 \\ \frac{C_3(h_2(\mu),h_1(\mu);h_4(\mu))}{1-C_1(h_2(\mu),h_1(\mu);h_4(\mu))} \\ -\frac{C_3(h_3(\mu),h_1(\mu);h_4(\mu))}{C_1(h_3(\mu),h_1(\mu);h_4(\mu))} \\ 1 \end{pmatrix} \tag{3.7}$$

is implied by

$$h_1(\mu) = c_1(\zeta)$$
$$h_2(\mu) - C(h_2(\mu), h_1(\mu); h_4(\mu)) = c_2(\zeta) \tag{3.8}$$
$$C(h_3(\mu), h_1(\mu); h_4(\mu)) = c_3(\zeta)$$
$$h_4(\mu) = \pi + c_4(\zeta),$$

where $c_i(\zeta)$ is an arbitrary one-dimensional function of ζ for $i = 1, 2, 3, 4$. For Step 3, upon setting $c_1(\zeta) = \zeta_1$, $c_2(\zeta) = \zeta_2$, $c_3(\zeta) = \zeta_3$ and $c_4(\zeta) = 0$, we have

$$\frac{\partial h(\mu)}{\partial \mu'} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{C_2(h_2(\mu),\zeta_1;\pi)}{1-C_1(h_2(\mu),\zeta_1;\pi)} & \frac{1}{1-C_1(h_2(\mu),\zeta_1;\pi)} & 0 & \frac{C_3(h_2(\mu),\zeta_1;\pi)}{1-C_1(h_2(\mu),\zeta_1;\pi)} \\ -\frac{C_2(h_3(\mu),\zeta_1;\pi)}{C_1(h_3(\mu),\zeta_1;\pi)} & 0 & \frac{1}{C_1(h_3(\mu),\zeta_1;\pi)} & -\frac{C_3(h_3(\mu),\zeta_1;\pi)}{C_1(h_3(\mu),\zeta_1;\pi)} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3.9}$$

being full rank. Thus, we have found a reparameterization function $h(\cdot)$ satisfying the conditions of Algorithm 3.1 though its explicit form will depend upon the functional form of the copula $C(\cdot)$. For example, if we use the Ali-Mikhail-Haq copula, defined for $u_1, u_2 \in [0, 1]$ and $\pi \in [-1, 1)$ by

$$C(u_1, u_2; \pi) = \frac{u_1 u_2}{1 - \pi(1 - u_1)(1 - u_2)}, \tag{3.10}$$

we obtain the following closed-form solution for $h(\cdot)$:

$$h(\mu) = \begin{pmatrix} \zeta_1 \\ \frac{-b(\mu)+\sqrt{b(\mu)^2-4a(\mu)c(\mu)}}{2a(\mu)} \\ \frac{\zeta_3(1-\pi+\pi\zeta_1)}{\zeta_1-\zeta_3\pi+\zeta_1\zeta_3\pi} \\ \pi \end{pmatrix}, \tag{3.11}$$

where $a(\mu) = \pi(1 - \zeta_1)$, $b(\mu) = (1 - \zeta_1)(1 - \pi - \pi\zeta_2)$ and $c(\mu) = \zeta_2[\pi(1 - \zeta_1) - 1]$.[15] For any

---

[15]As may be gleaned from this formula, the expression for $h_2(\mu)$ comes from solving a quadratic equation.

choice of copula, we can also express the new parameters as a function of the original ones as follows:

$$\mu = (\zeta_1, \zeta_2, \zeta_3, \pi) = h^{-1}(\boldsymbol{\zeta}, \boldsymbol{\pi}) = (\boldsymbol{\zeta}, \boldsymbol{\pi}_1 - C(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3), C(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3), \boldsymbol{\pi}_3). \qquad (3.12)$$

Examples 2.4, continued. *In this example, we again consider GMM estimation so that* $\bar{\boldsymbol{g}}_n(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \varphi(W_i, \boldsymbol{\theta})$, *where the moment function* $\varphi(w, \boldsymbol{\theta})$ *is given by* (2.7). *The sample Jacobian with respect to* $\boldsymbol{\mu}$ *is*

$$\frac{\partial \bar{\boldsymbol{g}}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}'} = -\frac{1}{n} \sum_{i=1}^{n} A(Y_{h,i}) \begin{bmatrix} \boldsymbol{\pi}_2 + \boldsymbol{\pi}_3 & \boldsymbol{\pi}_1 & \boldsymbol{\pi}_1 \\ -\boldsymbol{\pi}_2 & 1 - \boldsymbol{\pi}_1 & 0 \end{bmatrix}$$

when $\beta = 0$. *Since again* $r = d_{\boldsymbol{\mu}} - 1$ *so that* $d_\pi = 1$, *simplifying notation as in the previous examples, for Step 1 of Algorithm 3.1, we set* $m(\boldsymbol{\mu}) = (-\boldsymbol{\pi}_1(1 - \boldsymbol{\pi}_1), -\boldsymbol{\pi}_1 \boldsymbol{\pi}_2, \boldsymbol{\pi}_2 + \boldsymbol{\pi}_3(1 - \boldsymbol{\pi}_1))'$. *For Step 2, we need to find the general solution in* $h(\cdot)$ *to the following system of ODEs:*

$$\frac{\partial h(\mu)}{\partial \pi} = (-h_1(\mu)(1 - h_1(\mu)), -h_1(\mu)h_2(\mu), h_2(\mu) + h_3(\mu)(1 - h_1(\mu)))'.$$

Given its triangular structure, this system can be solved successively using standard single-equation ODE methods, starting with the $\partial h_1(\mu)/\partial \pi$ *equation, then the* $\partial h_2(\mu)/\partial \pi$ *equation, followed by the* $\partial h_3(\mu)/\partial \pi$ *equation. The general solution takes the form*

$$h(\mu) = \begin{pmatrix} [1 + c_1(\zeta)e^\pi]^{-1} \\ c_2(\zeta)[e^{-\pi} + c_1(\zeta)] \\ c_3(\zeta)[1 + c_1(\zeta)e^\pi] - c_2(\zeta)[e^{-\pi} + c_1(\zeta)] \end{pmatrix},$$

where $c_i(\zeta)$ *is an arbitrary function of* ζ *for* $i = 1, 2, 3$. *For Step 3, setting* $c_1(\zeta) = 1$, $c_2(\zeta) = e^{\zeta_1}$ *and* $c_3(\zeta) = \zeta_2$ *induces a simple triangular structure on the components of* $h(\mu)$ *as functions of* μ, *i.e., so that* $h_1(\mu)$ *is a function of* π *only and* $h_2(\mu)$ *is a function of* π *and* ζ_1 *only. Such a triangular structure makes it easier to solve for* μ *in terms of* $\boldsymbol{\mu}$. *In this case, we have*

$$\frac{\partial h(\mu)}{\partial \mu'} = \begin{bmatrix} 0 & 0 & -e^\pi(1 + e^\pi)^{-2} \\ e^{\zeta_1}(e^{-\pi} + 1) & 0 & -e^{\zeta_1 - \pi} \\ -e^{\zeta_1}(e^{-\pi} + 1) & 1 + e^\pi & \zeta_2 e^\pi + e^{\zeta_1 - \pi} \end{bmatrix}$$

being full rank. Thus, we have found a reparameterization function $h(\cdot)$ *satisfying the conditions*

This solution has two solutions, one of which is always negative and one of which is always positive. Given that $h_2(\mu) = \boldsymbol{\pi}_1$ must be positive, $h_2(\mu)$ is equal to the positive solution.

of Algorithm 3.1 such that $\boldsymbol{\mu} = h(\mu) = (1/(1 + e^\pi), e^{\zeta_1}(e^{-\pi} + 1), \zeta_2(1 + e^\pi) - e^{\zeta_1}(e^{-\pi} + 1))$, *or equivalently,* $\mu = (\zeta_1, \zeta_2, \pi) = (\log(\boldsymbol{\pi}_2(1 - \boldsymbol{\pi}_1)), \boldsymbol{\pi}_1(\boldsymbol{\pi}_2 + \boldsymbol{\pi}_3), \log((1 - \boldsymbol{\pi}_1)/\boldsymbol{\pi}_1))$.

4 Limit Theory for Extremum Estimators

We proceed to derive the limit theory for the extremum estimator $\hat{\boldsymbol{\theta}}_n$ under a comprehensive class of identification strengths by applying results from AC12 to the estimator of the parameters in the reparameterized model $\hat{\theta}_n$ and then determining the asymptotic behavior of the original parameter estimator of interest via the relation $\hat{\boldsymbol{\theta}}_n = (\hat{\beta}_n, h(\hat{\mu}_n))$. 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 (which includes non-identification). Our ultimate goal from deriving asymptotic theory under parameters with different strengths of identification is to conduct uniformly valid inference that is robust to identification strength.

The true parameter space $\boldsymbol{\Gamma}$ for $\boldsymbol{\gamma}$ takes the form

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

where $\boldsymbol{\Theta}^*$ is a compact subset of $\mathbb{R}^{d_{\boldsymbol{\theta}}}$ and $\boldsymbol{\Phi}^*(\boldsymbol{\theta}) \subset \Phi^*$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*$ 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 $\bar{h}(\theta) \equiv (\beta, h(\mu))$ where h is the solution from Procedure 3. The next lemma formally establishes the properties of the reparameterization function $\bar{h}(\cdot)$.

Assumption H. *(i) $h : \mathcal{M} \to \boldsymbol{\mathcal{M}}$ is proper and continuously differentiable; (ii) $\boldsymbol{\Theta}$ is simply connected.*

Sufficient conditions for Assumption H(i) are (i) $\mathcal{M}$ is bounded and (ii) h is continuously differentiable.[16]

Lemma 4.1. *Under Assumptions Reg1, ID and H, (i) the function $\bar{h} : \Theta \to \boldsymbol{\Theta}$ is a homeomorphism and hence bijective; (ii) $\bar{h}(\theta)$ is continuously differentiable on Θ.*

Lemma 4.1(i) implies the bijectivity of $\bar{h} : \Theta^* \to \boldsymbol{\Theta}^*$ as well, since we assume that the true parameter space is contained in the optimizing parameter space.[17] Due to this result, we can

---

[16]A function is proper if its pre-image of a compact set is compact. If h is continuous, the pre-image of a closed set under h is closed. Also, if $\mathcal{M}$ is bounded, the pre-image of a bounded set under h is bounded. Therefore, under these sufficient conditions, h is proper.

[17]See Assumption B1 in AC12, which is imposed in Theorem 4.1, Corollary 4.1 and Proposition 5.1 below.

equivalently derive limit theory derived under sequences of parameters in $\boldsymbol{\Gamma}$ or in the following transformed parameter space:

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

where $\Theta^* \equiv \bar{h}^{-1}(\boldsymbol{\Theta}^*)$ and $\Phi^*(\theta) \equiv \boldsymbol{\Phi}^*(\bar{h}(\theta)) \subset \Phi^*$ for all $\theta \in \Theta^*$.

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

$$\Gamma(\gamma_0) \equiv \{\{\gamma_n \in \Gamma : n \geq 1\} : \gamma_n \to \gamma_0 \in \Gamma\},$$

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

$$\Gamma(\gamma_0, \infty, \omega_0) \equiv \left\{\{\gamma_n\} \in \Gamma(\gamma_0) : n^{1/2} \|\beta_n\| \to \infty \text{ and } \frac{\beta_n}{\|\beta_n\|} \to \omega_0 \in \mathbb{R}^{d_\beta}\right\},$$

where $\gamma_0 \equiv (\theta_0, \phi_0)$ and $\gamma_n \equiv (\theta_n, \phi_n)$, 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 characterize the limit theory for subvectors of the original parameter estimator of interest $\hat{\boldsymbol{\theta}}_n$, which we show is equal to $\bar{h}(\hat{\theta}_n)$ by using Lemma 4.1. Toward this end, we use $\hat{\boldsymbol{\mu}}_n^s$ to denote a generic d_s-dimensional subvector of $\hat{\boldsymbol{\mu}}_n$ and $h^s(\cdot)$ to denote the corresponding elements of $h(\cdot)$ in the relation $\hat{\boldsymbol{\mu}}_n = h(\hat{\mu}_n)$. Let $h_\mu^s(\mu) = \partial h^s(\mu)/\partial \mu'$ and partition $h_\mu^s(\mu)$ conformably with $\mu = (\zeta, \pi)$: $h_\mu^s(\mu) = [h_\zeta^s(\mu) : h_\pi^s(\mu)]$. Suppose $rank(h_\pi^s(\mu)) = \tilde{d}_\pi^*$ for all $\mu \in \mathcal{M}_\epsilon \equiv \{\mu : (\beta, \mu) \in \Theta, \|\beta\| < \epsilon\}$ for some $\epsilon > 0$. For $\mu \in \mathcal{M}_\epsilon$, let $\tilde{A}(\mu) \equiv [\tilde{A}_1(\mu)' : \tilde{A}_2(\mu)']'$ be an orthogonal $d_s \times d_s$ matrix such that $\tilde{A}_1(\mu)$ is a $(d_s - \tilde{d}_\pi^*) \times d_s$ matrix whose rows span the null space of $h_\pi^s(\mu)'$ and $\tilde{A}_2(\mu)$ is a $\tilde{d}_\pi^* \times d_s$ matrix whose rows span the column space of $h_\pi^s(\mu)$. The matrix $\tilde{A}_1(\mu)$ essentially rotates $h^s(\mu)$ "off" the π direction of its parameter space while the matrix $A_2^s(\mu)$ rotates $h^s(\mu)$ "in" the direction of π. The estimate $\hat{\boldsymbol{\mu}}_n^s = h^s(\hat{\mu}_n)$ 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 related contexts 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), Sims et al. (1990), Antoine and Renault (2009, 2012), AC14 and Phillips (2016).

The following assumptions impose regularity conditions on the subvector function $h^s(\cdot)$.

Assumption Reg2. $rank(h_\pi^s(\mu)) = \tilde{d}_\pi^*$ *for some constant* $\tilde{d}_\pi^* \leq d_\pi$ *for all* $\mu \in \mathcal{M}_\epsilon$ *for some* $\epsilon > 0$.

Define

$$\tilde{\eta}_n(\mu) \equiv \begin{cases} \sqrt{n}\tilde{A}_1(\mu)\{h^s(\zeta_n, \pi) - h^s(\zeta_n, \pi_n)\}, & \text{if } \tilde{d}^*_\pi < d_s \\ 0, & \text{if } \tilde{d}^*_\pi = d_s. \end{cases}$$

Assumption Reg3. *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $\tilde{\eta}_n(\hat{\mu}_n) \xrightarrow{p} 0$ *for all* $b \in \mathbb{R}^{d_\beta}_\infty$.

Analogous assumptions can be found in, e.g., Assumptions R1 and R2 of AC14. With an explicit $h(\cdot)$ found e.g., by Algorithm 3.1, Assumption Reg2 is straightforward to verify. Assumption Reg3 is a high-level assumption that may be verified via any of the sufficient conditions given in Assumption Reg3* below.

Assumption Reg3*. *(i)* $\tilde{d}^*_\pi = d_s$.
    (ii) $d_s = 1$.
    (iii) The column space of $h^s_\pi(\mu)$ *is the same for all* $\mu \in \mathcal{M}_\epsilon$ *for some* $\epsilon > 0$.
    (iv) $h^s(\mu) = H^s\mu$, *where* H^s *is a* $d_s \times d_\mu$ *matrix with full row rank.*
    (v) No more than d_π *entries of* $h^s(\mu)$ *depend upon* π *and each* π-*dependent entry depends on a single different element of* π.

Applying results of Lemmas 5.1 and 5.2 of AC14 shows that any of the conditions of Assumption Reg3*(i)-(iv) is sufficient for Assumption Reg3 to hold. The condition in Assumption Reg3*(v) is sufficient for the condition in Assumption Reg3*(iii) to hold, as formalized in the following lemma. This condition is relevant when the reparameterization function $h(\cdot)$ is non-linear and one wishes to obtain the joint limiting behavior of a larger subvector of $\hat{\mu}_n$ such that $d_s > \max\{\tilde{d}^*_\pi, 1\}$. As may be gleaned from the sufficient conditions of Assumption Reg3*, the feasibility of rotating a subvector $\hat{\boldsymbol{\mu}}^s_n$ to obtain a $\sqrt{n}$-convergent direction in the parameter space requires restrictions on the number of entires of $\hat{\boldsymbol{\mu}}^s_n = h^s(\hat{\mu}_n)$ that are nonlinear functions of $\hat{\pi}_n$. These types of restrictions will be important for conducting Wald statistic-based inference in the next section and are explored in more detail in the context of Example 2.3 after the following lemma.

Lemma 4.2. *Assumption Reg3*(v) implies Assumption Reg3*(iii).*

Examples 2.3, continued. *We first note that by expression* (3.11), *Assumption Reg3*(v) holds for any two-dimensional subvector* $h^s(\mu) = (h_1(\mu), h_j(\mu))$ *for any* $j = 2, 3$ *or* 4. *Thus, we may rotate any corresponding* $\hat{\boldsymbol{\mu}}^s_n = (\hat{\boldsymbol{\mu}}_{n,1}, \hat{\boldsymbol{\mu}}_{n,j})$ *to find a* $\sqrt{n}$-*convergent direction of the parameter space and apply the limit theory of the following theorem, even for those* $\boldsymbol{\mu}_j$'s *that are nonlinear functions of* π *(i.e., for* $j = 2$ *or* 3*). On the other hand, none of the conditions of Assumption Reg3* hold for any* $\hat{\boldsymbol{\mu}}^s_n$ *containing more than one* $\hat{\boldsymbol{\mu}}_{n,j}$ *for* $j = 2, 3$ *or* 4 *and it is not possible to*

find a $\sqrt{n}$-convergent rotation. For illustration, consider the simplest of these cases for which $\hat{\boldsymbol{\mu}}_n^s = (\hat{\boldsymbol{\mu}}_{n,3}, \hat{\boldsymbol{\mu}}_{n,4})$. In this case under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$,

$$\tilde{A}_1(\hat{\mu}_n) = \mathcal{S}(\hat{\mu}_n) \left(1, \frac{C_3(h_3(\hat{\mu}_n), \hat{\zeta}_{1,n}; \hat{\pi}_n)}{C_1(h_3(\hat{\mu}_n), \hat{\zeta}_{1,n}; \hat{\pi}_n)} \right),$$

where $\mathcal{S}(\hat{\mu}_n) \equiv \{1 + C_3(h_3(\hat{\mu}_n), \hat{\zeta}_{1,n}; \hat{\pi}_n)^2 / C_1(h_3(\hat{\mu}_n), \hat{\zeta}_{1,n}; \hat{\pi}_n)^2\}^{-1/2}$ so that

$$\tilde{\eta}_n(\hat{\mu}_n) = \sqrt{n} \mathcal{S}(\hat{\mu}_n) \left[\frac{\tilde{\eta}_n^N(\hat{\mu}_n)}{\tilde{\eta}_n^D(\hat{\mu}_n)} \right] (\hat{\pi}_n - \pi_n),$$

where

$$\tilde{\eta}_n^N(\hat{\mu}_n) \equiv \zeta_{3,n}^2 (\zeta_{1,n} - 1)^2 (\zeta_{1,n} - \zeta_{3,n})(\hat{\pi}_n - \pi_n) + O_p(n^{-1/2}) = O_p(n^{-1/2} \|\beta_n\|^{-1}),$$
$$\tilde{\eta}_n^D(\hat{\mu}_n) \equiv \{\zeta_{1,n} - \zeta_{3,n} \hat{\pi}_n + \zeta_{1,n} \zeta_{3,n} \hat{\pi}_n + O_p(n^{-1/2})\}^2 (\zeta_{1,n} - \zeta_{3,n} \pi + \zeta_{1,n} \zeta_{3,n} \pi) = O_p(1),$$

and $\mathcal{S}(\hat{\mu}_n) = O_p(1)$, which we obtain by using the results from Lemma A.1 in Appendix A. (The derivations behind the above expressions can be found in Appendix B.) Thus, we have that $\|\tilde{\eta}_n(\hat{\mu}_n)\| = \|O_p(n^{-1/2} \|\beta_n\|^{-1}) \sqrt{n}(\hat{\pi}_n - \pi_n)\| = \|O_p(n^{-1/2} \|\beta_n\|^{-2})\| \to \infty$ if $n^{1/4} \|\beta_n\| \to 0$, according to Lemma A.1.

Define

$$\iota(\beta) \equiv \begin{cases} \beta, & \text{if } \beta \text{ is scalar,} \\ \|\beta\|, & \text{if } \beta \text{ is a vector.} \end{cases}$$

We are now ready to state the main result of this section.

Theorem 4.1. *(i) Suppose Assumptions CF, Reg1, ID, Jac, Reg2, Reg3 and H, and Assumptions B1-B3 and C1-C6 of AC12,[18] applied to the transformed objects of this paper including θ and $Q_n(\theta)$, hold. Under parameter sequences $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$,*

$$\begin{pmatrix} \sqrt{n}(\hat{\beta}_n - \beta_n) \\ \sqrt{n} \tilde{A}_1(\hat{\mu}_n)(\hat{\boldsymbol{\mu}}_n^s - \boldsymbol{\mu}_n^s) \\ \tilde{A}_2(\hat{\mu}_n)(\hat{\boldsymbol{\mu}}_n^s - \boldsymbol{\mu}_n^s) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tau_{0,b}^\beta(\pi_{0,b}^*) \\ \tilde{A}_1(\zeta_0, \pi_{0,b}^*) h_\zeta^s(\zeta_0, \pi_{0,b}^*) \tau_{0,b}^\zeta(\pi_{0,b}^*) \\ \tilde{A}_2(\zeta_0, \pi_{0,b}^*)[h^s(\zeta_0, \pi_{0,b}^*) - \boldsymbol{\mu}_0^s] \end{pmatrix},$$

where

$$\pi_{0,b}^* \equiv \pi^*(\gamma_0, b) \equiv \arg \min_{\pi \in \Pi} -\frac{1}{2}(G_0(\pi) + K_0(\pi)b)' H_0^{-1}(\pi)(G_0(\pi) + K_0(\pi)b),$$

---

[18]Here and below, we refer the reader to AC12 for the assumptions in that paper. For the sake of brevity, we do not repeat them in the current paper. In Appendix B, however, we provide sufficient conditions for all the assumptions used in this paper including those from AC12 for the threshold crossing model (Example 2.3).

$$\tau_{0,b}(\pi) \equiv \tau(\pi; \gamma_0, b) \equiv -H_0^{-1}(\pi)(G_0(\pi) + K_0(\pi)b) - (b, 0_{d_\zeta \times 1})$$

with $\pi_{0,b}^*$ being a random vector that minimizes a non-central chi-squared process and $\{\tau_{0,b}(\pi) : \pi \in \Pi\}$ being a Gaussian process for which $\tau_{0,b}^\beta(\pi)$ and $\tau_{0,b}^\zeta(\pi)$ denote the first d_β and final $d_\mu - d_\pi$ entries. The underlying Gaussian process $G_0(\cdot) \equiv G(\cdot; \gamma_0)$ is defined in Assumption C3 of AC12 and 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 CF, Reg1, ID, Jac, Reg2, Reg3 and H, 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 $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

$$\sqrt{n} \begin{pmatrix} \hat{\beta}_n - \beta_n \\ \tilde{A}_1(\hat{\mu}_n)(\hat{\mu}_n^s - \mu_n^s) \\ \iota(\beta_n)\tilde{A}_2(\hat{\mu}_n)(\hat{\mu}_n^s - \mu_n^s) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z_\beta \\ \tilde{A}_1(\mu_0)h_\zeta^s(\mu_0)Z_\zeta \\ \tilde{A}_2(\mu_0)h_\pi^s(\mu_0)Z_\pi \end{pmatrix},$$

if $\beta_0 = 0$ and

$$\sqrt{n} \begin{pmatrix} \hat{\beta}_n - \beta_n \\ \hat{\mu}_n - \mu_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z_\beta \\ h_\zeta(\mu_0)Z_\zeta + \iota(\beta_0)^{-1}h_\pi(\mu_0)Z_\pi \end{pmatrix}$$

if $\beta_0 \neq 0$, where $(Z_\beta, Z_\zeta, Z_\pi) = Z_\theta \sim \mathcal{N}(0, J^{-1}(\gamma_0)V(\gamma_0)J^{-1}(\gamma_0))$. The underlying matrices $J(\gamma_0)$ and $V(\gamma_0)$ are defined in Assumptions D2 and D3 of AC12.

Theorem 4.1 describes the joint limiting behavior of $\hat{\beta}_n$ and $\hat{\mu}_n^s$ under a comprehensive class of identification strengths. By rotating the subvector $\hat{\mu}_n^s$ in the appropriate direction of the parameter space via $A_1(\hat{\mu}_n)$, we obtain $\sqrt{n}$-consistency under weak and semi-strong identification. If the full vector function $h(\cdot)$ satisfies Assumptions Reg2 and Reg3, then the results of Theorem 4.1 apply to the full parameter vector $\hat{\mu}_n$. Though nonlinearity of the reparameterization function often makes it impossible to obtain a $\sqrt{n}$-consistent rotation of the full vector $\hat{\mu}_n$ under weak and semi-strong identification, it is still possible to characterize its joint limiting behavior at slower convergence rates without rotation, as in the following corollary.

In order to express this corollary, it is necessary to separate the components of $\mu = h(\zeta, \pi)$ according to whether they depend upon π or not. Without loss of generality, suppose that the first d_{μ^1} components of $h(\zeta, \pi)$ do not actually depend upon π (e.g., in cases described by Remark 3.5), while the final $d_\mu - d_{\mu^1}$ of $h(\zeta, \pi)$ do. Denote the corresponding entries of $\mu = h(\zeta, \pi)$ as $\mu^1 = h^1(\zeta)$ and $\mu^2 = h^2(\zeta, \pi)$, respectively.

Corollary 4.1. *Suppose all of the assumptions of Theorem 4.1 hold except for Assumption Reg3. Under parameter sequences $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$,*

(i)

$$\begin{pmatrix} \sqrt{n}(\hat{\beta}_n - \beta_n) \\ \sqrt{n}(\hat{\boldsymbol{\mu}}_n^1 - \boldsymbol{\mu}_n^1) \\ \hat{\boldsymbol{\mu}}_n^2 \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \tau_{0,b}^{\beta}(\pi_{0,b}^*) \\ h_{\zeta}^1(\zeta_0)\tau_{0,b}^{\zeta}(\pi_{0,b}^*) \\ h^2(\zeta_0, \pi_{0,b}^*) \end{pmatrix}$$

if $\|b\| < \infty$ *and*

(ii)

$$\sqrt{n} \begin{pmatrix} \hat{\beta}_n - \beta_n \\ \hat{\boldsymbol{\mu}}_n^1 - \boldsymbol{\mu}_n^1 \\ \iota(\beta_n)(\hat{\boldsymbol{\mu}}_n^2 - \boldsymbol{\mu}_n^2) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z_{\beta} \\ h_{\zeta}^1(\zeta_0)Z_{\zeta} \\ h_{\pi}^2(\mu_0)Z_{\pi} \end{pmatrix},$$

if $\|b\| = \infty$.

Apart from the simpler cases for which $d_{\boldsymbol{\mu}^2} = d_{\pi}$ that are already covered by the analysis of AC12, it is interesting to note that the limiting random vectors under both cases of Corollary 4.1 are singular in some sense. For case (ii), the singularity is straightforward: the Gaussian limit has a singular covariance matrix. For case (i), the singularity comes from the fact that $dim(\pi_{0,b}^*) = d_{\pi} < d_{\boldsymbol{\mu}^2} = d_{\boldsymbol{\mu}} - d_{\boldsymbol{\mu}^1}$ so that the dimension of the parameter estimator $\hat{\boldsymbol{\mu}}_n^2$ exceeds the dimension of the "randomness" in its limit.

5 Wald Statistics

We are interested in testing general nonlinear hypotheses of the form

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

using the Wald statistic. To reduce notation and make assumptions more transparent, it is useful to view H_0 in its equivalent form as a hypothesis on the reparameterized parameters θ, viz.,

$$H_0 : r(\theta) \equiv \boldsymbol{r}(\bar{h}(\theta)) = v \in \mathbb{R}^{d_r},$$

With this notation in mind, a standard Wald statistic for H_0 based upon $\hat{\boldsymbol{\theta}}_n = \bar{h}(\hat{\theta}_n)$ can be written as[19]

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

---

[19]The Wald statistic $W_n(v)$ is identical to the usual Wald statistic written as a function of $\hat{\boldsymbol{\theta}}_n$ that uses an estimator of the asymptotic covariance matrix for $\hat{\boldsymbol{\theta}}_n$ that takes the natural form $\bar{h}_{\theta}(\hat{\theta}_n)B^{-1}(\hat{\beta}_n)\hat{\Sigma}_n B^{-1}(\hat{\beta}_n)\bar{h}_{\theta}(\hat{\theta}_n)'$.

where $r_\theta(\theta) \equiv \partial r(\theta)/\partial \theta' \equiv [r_\beta(\theta) : r_\zeta(\theta) : r_\pi(\theta)] \in \mathbb{R}^{d_r \times d_\theta}$, $\hat\Sigma_n$ estimates the covariance matrix of $(Z'_\beta, Z'_\zeta, Z'_\pi)'$ and

$$B(\beta) = \begin{pmatrix} I_{d_\beta} & 0 & 0 \\ 0 & I_{d_\zeta} & 0 \\ 0 & 0 & \iota(\beta)I_{d_\pi} \end{pmatrix}.$$

Note that, although the asymptotic distributions we obtain under weak identification are not pivotal, scaling by $\hat\Sigma_n$ in the Wald statistic can still be motivated by asymptotic pivotality under (semi-)strong identification (see Proposition 5.1(ii)).

Under the assumptions of Theorem 4.1 and R1–R2 and V1–V2 of AC14, the limiting behavior of $W_n(v)$ under $\{\gamma_n\} \in \Gamma(\gamma_0, b)$ or $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ can be obtained as a simple application of the results of Theorem 5.1 of that paper. However, the fact that $\hat{\boldsymbol{\theta}}_n$ is generally a nonlinear function of $\hat\theta_n$ creates certain peculiarities specific to the current context of potential under-identification that are worth exploring in more detail. In particular, Assumptions R1 and R2 of AC14 rule out a handful of very standard null hypotheses that the Wald statistic can be used for in the presence of (near-)under-identification. Hence, we repeat these assumptions here and discuss them in the present context.

Assumption R1. (i) $r(\theta)$ is continuously differentiable on Θ.

(ii) $r_\theta(\theta)$ is full row rank d_r for all $\theta \in \Theta$.

(iii) $rank(r_\pi(\theta)) = d_\pi^*$ for some constant $d_\pi^* \leq min\{d_r, d_\pi\}$ for all $\theta \in \Theta_\epsilon \equiv \{\theta \in \Theta : \|\beta\| < \epsilon\}$ for some $\epsilon > 0$.

Assumption R1(i) holds in the present context if the restriction on the original parameters $\boldsymbol{r}(\boldsymbol{\theta})$ is continuously differentiable on $\boldsymbol{\Theta}$ because $\bar{h}(\theta)$ is continuously differentiable on Θ by Lemma 4.1(ii). Since $\bar{h}_\theta(\theta)$ is full rank by Lemma 4.1(i), Assumption R1(ii) holds if $\partial \boldsymbol{r}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}'$ is full row rank for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Finally, Assumption R1(iii) requires the product of $\partial \boldsymbol{r}(\bar{h}(\theta))/\partial \boldsymbol{\mu}'$ and $h_\pi(\theta)$ to have constant rank for all $\theta \in \Theta_\epsilon$, which should occur when they each separately have constant rank in the absence of some perverse interaction between them.

Let $A(\theta) = [A_1(\theta)' : A_2(\theta)']'$ be an orthogonal $d_r \times d_r$ matrix such that $A_1(\theta)$ is a $(d_r - d_\pi^*) \times d_r$ matrix whose rows span the null space of $r_\pi(\theta)'$ and $A_2(\theta)$ is a $d_\pi^* \times d_r$ matrix whose rows span the column space of $r_\pi(\theta)$. Let

$$\eta_n(\theta) \equiv \begin{cases} n^{1/2} A_1(\theta) \{r(\beta_n, \zeta_n, \pi) - r(\beta_n, \zeta_n, \pi_n)\}, & \text{if } d_\pi^* < d_r \\ 0, & \text{if } d_\pi^* = d_r. \end{cases}$$

Assumption R2. Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$, $\eta_n(\hat\theta_n) \xrightarrow{p} 0$ for all $b \in \mathbb{R}_\infty^{d_\beta}$.

In leading cases of interest, subvector null hypotheses, i.e., $H_0 : \boldsymbol{\theta}^s = v$ for some subvector $\boldsymbol{\theta}^s$ of $\boldsymbol{\theta}$, Assumption R2 is equivalent to Assumption Reg3 introduced in the previous section.[20] Recalling that Assumption Reg3 is used to show a $\sqrt{n}$-convergent rotation of $\hat{\boldsymbol{\theta}}_n^s$ can be constructed, we note that *the existence of such a $\sqrt{n}$-convergent rotation is crucial* to obtaining the convergence of a subvector Wald statistic under weak and semi-strong identification sequences. In the potential presence of the more complicated forms of identification failure we are interested in here, standard Wald statistics for testing seemingly straightforward (linear) hypotheses can easily diverge under the null hypothesis and weak or semi-strong identification sequences.

Remark 5.1. *In cases for which $\|\eta_n(\hat{\theta}_n)\|$ diverges, Theorem 5.2 of AC14 tells us that $W_n(v)$ also diverges. This is particularly important in the context of the nonlinear reparameterizations of this paper. For example, it implies that if the reparameterization function $h(\cdot)$ is nonlinear, a standard subvector Wald statistic can easily diverge when the subvector under test is "large enough", containing more than d_π entries of $\boldsymbol{\mu}$ that are nonlinear functions of π. See the continuation of Example 2.3 in the previous section for an example. This result is very important in practice. It implies that subvector Wald tests making use of $\chi^2_{d_r}$ CVs exhibit size distortion of the most extreme kind: their asymptotic size is equal to one if the subvector is large enough (including the full vector $\boldsymbol{\theta}$).*

Any one of the following sufficient conditions implies the high-level Assumption R2, as verified in Lemma 5.1 of AC14.

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

In our context, Assumption R2*(i) requires the number of restrictions under test not exceed d_π and that all restrictions must involve elements of $\boldsymbol{\mu}$ that are nontrivial functions of π. In the case of subvector hypotheses, Assumption R2*(i)-(iii) is identical to Assumption Reg3*(i)-(iii) and Assumptions Reg3*(iv) and (v) each implies Assumption R2*(iii).[21]

Assumption R$_\text{L}$. *$r(\theta) = R\theta$, where R is a $d_r \times d_\theta$ matrix with full row rank.*

In the present context, Assumption R$_\text{L}$ essentially requires both the reparameterization function $h(\cdot)$ and the restrictions under test to be linear, viz., $h(\theta) = H\theta$ and $\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{R\theta}$ so that $r(\theta) = \boldsymbol{R}H\theta$. The reparameterization function $h(\cdot)$ is not generally linear. However, it is

---

[20]This statement holds because if any elements of $\boldsymbol{r}(\boldsymbol{\theta})$ are equal to elements of β, the corresponding elements of $r(\beta_n, \zeta_n, \pi) - r(\beta_n, \zeta_n, \pi_n)$ are simply equal to zero.
[21]These statements hold because β is not a function of π.

sometimes possible to obtain linear reparameterizations in special cases for which the underlying model is linear. See Remark 3.3. In linear models for which $h(\theta) = H\theta$, the Wald statistic for linear restrictions does not diverge under weak or semi-strong identification. The potential for Wald statistic divergence for linear (including subvector) restrictions under weak or semi-strong identification, as discussed in Remark 5.1, is truly a consequence of the nonlinearity of the models we study in this paper.

Under a sequence $\{\gamma_n\}$, we consider the sequence of null hypotheses $H_0 : \boldsymbol{r}(\boldsymbol{\theta}) = v_n$, where $v_n = \boldsymbol{r}(\boldsymbol{\theta}_n)$. In combination with our reparameterization results, direct application of Theorem 5.1 of AC14 yields the following results.

Proposition 5.1. *(i) Suppose Assumptions CF, ID, Reg1, Jac, H, R1 and R2, 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 $\|b\| < \infty$,*

$$W_n(v_n) \xrightarrow{d} \lambda(\pi_{0,b}^*; \gamma_0, b),$$

where $\{\lambda(\pi; \gamma_0, b) : \pi \in \Pi\}$ is a stochastic process defined in expression (5.20) of AC14.

(ii) Suppose Assumptions CF, ID, Reg1, Jac, H, R1 and R2, 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 $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

$$W_n(v_n) \xrightarrow{d} \chi_{d_{\boldsymbol{r}}}^2.$$

Remark 5.2. *For some hypotheses, one may use the Wald statistic and robust CVs described in the following section to conduct tests that uniformly control asymptotic size in the potential presence of general identification failure. To better fit this result into the current literature on hypothesis testing that is robust to general forms of identification failure, we remark here on three leading categories of hypotheses that are of typical interest in applied work: (i) one-dimensional hypotheses, (ii) subvector hypotheses and (iii) full vector hypotheses. Our results are the first we are aware of that allow one to directly conduct one-dimensional hypothesis tests for general moment condition or likelihood models that fall into the framework of this paper. The methods of Andrews and Mikusheva (2016b) can only be used for these cases when the estimation problem can be formulated in a MD framework. To use the methods of Andrews and Guggenberger (2014) and Andrews and Mikusheva (2016a), one must rely on a power-reducing projection or Bonferroni bound-based approach. For subvector hypotheses, our results allow one to directly conduct hypothesis tests for a class of subvectors that are typically not "too large" (see Example 2.3 in Section 4 and Remark 5.1). On the other hand, one may "concentrate out" well-identified parameters to directly conduct hypothesis tests for a different class of subvectors in moment condition models using the methods of Andrews and Guggenberger (2014) and Andrews*

and Mikusheva (2016a).[22] *There is an interesting complementarity here between our results and those of Andrews and Guggenberger (2014) and Andrews and Mikusheva (2016a): to use the approach of these latter papers, the subvector must contain all parameters subject to identification failure so that, in some sense, the subvectors cannot be "too small". Finally, we note that except for models that already fall under the framework of AC12, the results of our paper do not allow one to directly conduct full vector hypotheses (due to the divergence of $\eta_n(\hat{\theta}_n)$) whereas the methods of Andrews and Guggenberger (2014) and Andrews and Mikusheva (2016a) do. We should also note that the frameworks of our paper and Andrews and Guggenberger (2014) or Andrews and Mikusheva (2016a) are non-nested.*

Remark 5.3. *We restrict focus in this paper to Wald statistics (rather than e.g., Langrange multiplier or likelihood ratio statistics) since they do not require estimation under the null hypothesis. This allows us to use the results of Section 4 and avoid restrictive assumptions on the reparameterization function $h(\cdot)$ and/or the restrictions under test $\boldsymbol{r}(\cdot)$. For example, AC12 impose Assumption RQ1(iii) to analyze the likelihood ratio statistic. Though somewhat restrictive even in their setting, such an assumption would be especially restrictive in our's since it would typically require the separate elements of $h(\cdot)$ to be functions of ζ or π only, but not both at the same time.*

6 Robust Wald Inference

The limit distribution of $\lambda(\pi_{0,b}^*; \gamma_0, b)$ given in Proposition 5.1(i) provides a good approximation to the finite-sample distribution of $W_n(v)$. 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 $\lambda(\pi_{0,b}^*; \gamma_0, b)$ in Proposition 5.1(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 ϕ so we can write $\phi = (\delta, \varphi)$.[23]

---

[22]Andrews and Mikusheva (2016a) cannot handle moment conditions for which the asymptotic variance matrix of the moments is singular. This occurs for the ML estimators of this paper.

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

Assumption FD. *The distribution function of $\lambda(\pi_{0,b}^*; \gamma_0, b)$ depends upon γ_0 only through ζ_0, π_0, and some $\delta_0 \in \mathbb{R}_\infty^{d_\delta}$ 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 $\hat{\delta}_n$ with $\hat{\delta}_n \xrightarrow{p} \delta_0$.*

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

6.1 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 $t_n \equiv (n\hat{\beta}_n' \hat{\Sigma}_{\beta\beta,n}^{-1} \hat{\beta}_n / d_\beta)^{1/2}$, where $\hat{\Sigma}_{\beta\beta,n}$ is equal to the upper left $d_\beta \times d_\beta$ block of $\hat{\Sigma}_n$ 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:

$$
c_{1-\alpha,n}^{ICS} \equiv \begin{cases} \chi_{d_r}^2 (1-\alpha)^{-1} & \text{if } t_n > \kappa_n, \\ c_{1-\alpha,n}^{LF} & \text{if } t_n \leq \kappa_n \end{cases}
$$

where $\chi_{d_r}^2 (1-\alpha)^{-1}$ is the $(1-\alpha)$ quantile of a $\chi_{d_r}^2$-distributed random variable and $c_{1-\alpha,n}^{LF} \equiv \sup_{\ell \in \hat{\mathcal{L}}_n \cap \mathcal{L}(v)} c_{1-\alpha}(\ell)$ with $\hat{\mathcal{L}}_n \equiv \{\ell = (b, \gamma) \in \mathcal{L} : \gamma = (\beta, \hat{\zeta}_n, \pi, \hat{\delta}_n, \varphi)\}$, $\mathcal{L}(v) \equiv \{\ell = (b, \gamma_0) \in \mathcal{L} : r(\theta_0) = v\}$, and $\mathcal{L} \equiv \{\ell = (b, \gamma_0) \in \mathbb{R}_\infty^{d_\beta} \times \Gamma : \text{for some } \{\gamma_n\} \in \Gamma(\gamma_0), n^{1/2}\beta_n \to b\}$. That is, we both impose H_0 and "plug-in" consistent estimators $\hat{\zeta}_n$ and $\hat{\delta}_n$ 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).[24] A typical choice for κ_n is $\kappa_n = (\log n)^{1/2}$ as it is analogous to the penalty term in the Bayesian information criterion. Under the assumptions of Proposition 5.1, Assumption FD and the following assumption, we can establish the correct asymptotic size of tests using the Wald statistic and ICS CVs.

Assumption DF1. *The distribution function of $\lambda(\pi_{0,b}^*; \gamma_0, b)$ is continuous at $\chi_{d_r}^2 (1-\alpha)^{-1}$ and $\sup_{\ell \in \mathcal{L}_0 \cap \mathcal{L}(v)} c_{1-\alpha}(\ell)$, where $\mathcal{L}_0 \equiv \{\ell = (b, \gamma) \in \mathcal{L} : \gamma = (\beta, \zeta_0, \pi, \delta_0, \varphi)\}$.*

This assumption is assured to hold e.g., if the distribution function of $\lambda(\pi_{0,b}^*; \gamma_0, b)$ is absolutely continuous. This both holds and is easy to check in most examples.

Proposition 6.1. *Under the assumptions of Proposition 5.1, Assumption K of AC12 and Assumptions FD and DF1, $\limsup_{n \to \infty} \sup_{\gamma \in \Gamma : r(\theta) = v} P_\gamma(W_n(v) > c_{1-\alpha,n}^{ICS}) = \alpha$.*

---

[24]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 $\hat{\mathcal{L}}_n \cap \mathcal{L}(v)$ with $\hat{\mathcal{L}}_n$ in the expression for $c_{1-\alpha}^{LF}$. Also, any consistent estimators of the components of γ_0 may be analogously "plugged-in".

6.2 Adjusted-Bonferroni CVs

The second type of CV we consider is a modification of the adjusted-Bonferroni CV of McCloskey (2017). 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 $\mathcal{P}(\hat{\zeta}_n, \hat{\delta}_n) \equiv \{(b,\pi) \in \mathbb{R}_\infty^{d_\beta + d_\pi} : \text{ for some } \gamma_0 \in \Gamma \text{ with } \zeta_0 = \hat{\zeta}_n \text{ and } \delta_0 = \hat{\delta}_n, \ \pi = \pi_0 \text{ and for some } \{\gamma_n\} \in \Gamma(\gamma_0), n^{1/2}\beta_n \to b\}$, as in the construction of least-favorable CVs, to a data-dependent set. Then one subsequently maximizes $c_{1-\alpha}(b,\gamma)$ over b and π in this restricted set. Intuitively, 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 $\hat{b}_n = n^{1/2}\hat{\beta}_n$. Using the results of Theorem 4.1, we can determine the joint asymptotic distribution of $(\hat{b}_n, \hat{\pi}_n)$ 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 (2017) uses such a confidence set for (b, π_0) as the data-dependent set to maximize $c_{1-\alpha}(b,\gamma)$ over. 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 $(\tau_{0,b}^\beta(\pi_{0,b}^*), \pi_{0,b}^*)$ depend upon the underlying parameters (b, π_0) themselves.[25] As a modification, here we instead propose the use of either one of two sets as follows. For notational simplicity, we will denote either of the two sets as $\hat{I}_n^a(\hat{b}_n, \hat{\pi}_n)$, though the second one does not depend directly on $\hat{\pi}_n$. The first is
$\hat{I}_n^a(\hat{b}_n, \hat{\pi}_n) = \{(b,\pi) \in \mathcal{P}(\hat{\zeta}_n, \hat{\delta}_n) : [(\hat{b}_n - b)', (\hat{\pi}_n - \pi)']\hat{\bar{\Sigma}}_n^{-1}[(\hat{b}_n - b)', (\hat{\pi}_n - \pi)']' \leq \chi^2_{d_\beta + d_\pi}(1-a)^{-1}\}$,
where
$$\hat{\bar{\Sigma}}_n \equiv \begin{pmatrix} \hat{\Sigma}_{\beta\beta,n} & n^{-1/2}\|\hat{\beta}_n\|^{-1}\hat{\Sigma}_{\beta\pi,n} \\ n^{-1/2}\|\hat{\beta}_n\|^{-1}\hat{\Sigma}'_{\beta\pi,n} & n^{-1}\|\hat{\beta}_n\|^{-2}\hat{\Sigma}_{\pi\pi,n} \end{pmatrix}$$

with $\hat{\Sigma}_{\beta\pi,n}$ denoting the upper right $d_\beta \times d_\pi$ block of $\hat{\Sigma}_n$ and $\hat{\Sigma}_{\pi\pi,n}$ denoting the lower right $d_\pi \times d_\pi$ block of $\hat{\Sigma}_n$. This set is akin to an a-level Wald confidence set for (b, π_0). The second set we propose can ease later computations: $\hat{I}_n^a(\hat{b}_n, \hat{\pi}_n) = \{(b,\pi) \in \mathcal{P}(\hat{\zeta}_n, \hat{\delta}_n) : (\hat{b}_n - b)'\hat{\Sigma}_{\beta\beta,n}^{-1}(\hat{b}_n - b) \leq \chi^2_{d_\beta}(1-a)^{-1}\}$. Though neither of these confidence sets has asymptotically correct coverage (at level $1-a$) under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$ sequences, they attain nearly correct coverage as $\|b\| \to \infty$. Similarly to the ICS CV in the previous subsection, one may also impose H_0 and "plug-in" the values of $\hat{\zeta}_n$ and $\hat{\delta}_n$ since they are consistent estimators.

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

---

[25]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.

1. Compute the smallest value $\varsigma = \varsigma(\hat{\zeta}_n, \hat{\delta}_n, \hat{\bar{\Sigma}}_n)$ such that

$$P\left(\lambda(\pi_{0,b}^*; \gamma_0, b) \geq \sup_{\ell \in \tilde{\mathcal{L}}_n^a(b,\gamma_0) \cap \mathcal{L}(v)} c_{1-\alpha}(\ell) + \varsigma\right) \leq \alpha$$

for all $(b, \gamma_0) \in \hat{\mathcal{L}}_n \cap \mathcal{L}(v)$.

2. Construct the quantity $c_{1-\alpha,n}^{AB} = \sup_{\ell \in \hat{\mathcal{L}}_n^a \cap \mathcal{L}(v)} c_{1-\alpha}(\ell) + \varsigma(\hat{\zeta}_n, \hat{\delta}_n, \hat{\bar{\Sigma}}_n)$. This is the adjusted-Bonferroni CV.

The computations in Step 1 can be achieved by simulating from the joint distribution of $\lambda(\pi_{0,b}^*; \gamma_0, b)$, $\tau_{0,b}^\beta(\pi_{0,b}^*)$ and $\pi_{0,b}^*$ over a grid of (b, γ_0) values in $\hat{\mathcal{L}}_n \cap \mathcal{L}(v)$ or by using more computationally efficient global optimization methods such as response surface analysis (see e.g., Jones et al., 1998 and Jones, 2001). See Algorithm Bonf-Adj in McCloskey (2017) for additional details on the computation of this CV. Under the assumptions of Proposition 5.1, Assumption FD and the following assumption, we can establish the correct asymptotic size of tests using the Wald statistic and adjusted-Bonferroni CVs.

Let $\mathcal{L}_0^a(b, \gamma_0) = \{\ell = (\tilde{b}, \gamma) \in \mathcal{L}_{\gamma_0} : (\tilde{b}, \pi) \in I_0^a(b + \tau_{0,b}^\beta(\pi_{0,b}^*), \pi_{0,b}^*)\}$, where $\mathcal{L}_{\gamma_0} \equiv \{\ell = (b, \gamma) \in \mathcal{L} : \gamma = (\beta, \zeta_0, \pi, \delta_0, \varphi)\}$. When using the first $\hat{I}_n^a(\hat{b}_n, \hat{\pi}_n)$ described above,

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

with

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

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 $\Sigma_0(\theta_{0,b}^*)$. (The function $\Sigma_0(\cdot)$ is defined in Assumptions V1 of AC12 and AC14.) When using the second $\hat{I}_n^a(\hat{b}_n, \hat{\pi}_n)$ described above,

$$I_0^a(b + \tau_{0,b}^\beta(\pi_{0,b}^*), \pi_{0,b}^*) = \{(b, \pi) \in \mathcal{P}(\zeta_0, \delta_0) : \tau_{0,b}^\beta(\pi_{0,b}^*)'\Sigma_{\beta\beta,0}^{-1}(\theta_{0,b}^*)\tau_{0,b}^\beta(\pi_{0,b}^*) \leq \chi_{d_\beta}^2(1 - a)^{-1}\}.$$

Assumption DF2. *There exists some* $(b^*, \gamma_0^*) \in \mathcal{L}$ *such that*

(i) $P(\lambda(\pi_{0,b^*}^*; \gamma_0^*, b^*) \geq \sup_{\ell \in \mathcal{L}_0^a(b^*, \gamma_0^*) \cap \mathcal{L}(v)} c_{1-\alpha}(\ell) + \varsigma(\zeta_0^*, \delta_0^*, \bar{\Sigma}(b^*, \gamma_0^*))) = \alpha,$

(ii) $P(\lambda(\pi_{0,b^*}^*; \gamma_0^*, b^*) = \sup_{\ell \in \mathcal{L}_0^a(b^*, \gamma_0^*) \cap \mathcal{L}(v)} c_{1-\alpha}(\ell) + \varsigma(\zeta_0^*, \delta_0^*, \bar{\Sigma}(b^*, \gamma_0^*))) = 0.$

This assumption is a similar distributional continuity condition to Assumption DF1 that holds in most examples.

Proposition 6.2. *Under the assumptions of Proposition 5.1 and Assumptions FD and DF2,* $\limsup_{n\to\infty} \sup_{\gamma\in\Gamma:r(\theta)=v} P_\gamma(W_n(v) > c_{1-\alpha,n}^{AB}) = \alpha.$

7 Threshold-Crossing Model Example

To illustrate our approach, we examine the threshold crossing model of a triangular system (Example 2.3) in this section. Weak identification and robust inference has been extensively studied in the literature (e.g., Staiger and Stock, 1997; Kleibergen, 2002; Moreira, 2003) for linear models of a triangular system (i.e, linear IV models), but not in this nonlinear setting. The latter, however, is empirically relevant when the dependent variable and endogenous regressor are both binary (e.g., Evans and Schwab, 1995; Goldman et al., 2001; Lochner and Moretti, 2004; Altonji et al., 2005; Rhine et al., 2006) and instruments are potentially weak.

The random sample 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 $\phi_0 \equiv \phi_{z,0} \equiv P_{\gamma_0}(Z_i = z)$. The ML estimator $\hat{\boldsymbol{\theta}}_n$ minimizes the following criterion function in $\boldsymbol{\theta} = (\beta, \boldsymbol{\zeta}, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \pi)$ over the parameter space $\boldsymbol{\Theta} \equiv \{\boldsymbol{\theta} = (\beta, \boldsymbol{\zeta}, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \pi) \in [-0.98 - \epsilon, 0.98 + \epsilon] \times [0.01 - \epsilon, 0.99 + \epsilon] \times [0.01 - \epsilon, 0.99 + \epsilon] \times [0.01 - \epsilon, 0.99 + \epsilon] \times [-0.99 - \epsilon, 0.99 + \epsilon] : 0.01 - \epsilon \le \beta + \boldsymbol{\zeta} \le 0.99 + \epsilon\}$:

$$\boldsymbol{Q}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\rho}(W_i, \boldsymbol{\theta})$$

for $\epsilon = 0.005$, where $\boldsymbol{\rho}(w, \boldsymbol{\theta}) \equiv -\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(w) \log \boldsymbol{p}_{yd,z}(\boldsymbol{\theta})$ is the logarithm of density function[26] with $\mathbf{1}_{ydz}(w) \equiv \mathbf{1}\{w = (y, d, z)\}$, and the set of $\boldsymbol{p}_{yd,z}(\boldsymbol{\theta})$'s are defined in (2.5)–(2.6).

7.1 Asymptotic Distributional Approximations for the Estimators

In this subsection, we describe the quantities composing the asymptotic distributions of the estimators in the threshold-crossing model example under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$ found in Theorem 4.1 and Corollary 4.1. The derivations used to obtain these quantities are given in Appendix B.

After the transformation, the transformed fitted probabilities $p_{yd,z}(\theta) \equiv \boldsymbol{p}_{yd,z}(\bar{h}(\theta))$ can be

---

[26]The log density would originally be $\boldsymbol{\rho}(w, \boldsymbol{\theta}, \phi) \equiv \sum_{y,d,z=0,1} \mathbf{1}_{ydz}(w)\{\log \boldsymbol{p}_{yd,z}(\boldsymbol{\theta}) + \log \phi_z\}$, but the term $\log \phi_z$ is dropped since it does not affect the optimization problem.

expressed as

$$
\begin{aligned}
p_{11,0}(\theta) &= \zeta_3, \\
p_{11,1}(\theta) &= C(h_3(\zeta_1, \zeta_3, \pi), \zeta_1 + \beta; \pi), \\
p_{10,0}(\theta) &= \zeta_2, \\
p_{10,1}(\theta) &= h_2(\zeta_1, \zeta_2, \pi) - C(h_2(\zeta_1, \zeta_2, \pi), \zeta_1 + \beta; \pi), \\
p_{01,0}(\theta) &= \zeta_1 - \zeta_3, \\
p_{01,1}(\theta) &= \zeta_1 + \beta - p_{11,1}(\theta),
\end{aligned}
\tag{7.1}
$$

and

$$
\begin{aligned}
p_{00,0}(\theta) &= 1 - p_{11,0}(\theta) - p_{10,0}(\theta) - p_{01,0}(\theta) = 1 - \zeta_1 - \zeta_2, \\
p_{00,1}(\theta) &= 1 - p_{11,1}(\theta) - p_{10,1}(\theta) - p_{01,1}(\theta) = 1 - \zeta_1 - \beta - p_{10,1}(\theta).
\end{aligned}
\tag{7.2}
$$

The first deterministic function appearing in the results of Theorem 4.1 and Corollary 4.1 is

$$
H(\pi; \gamma_0) = - \sum_{y,d,z=0,1} \frac{\phi_{z,0}}{p_{yd,z}(\theta_0)} D_\psi p_{yd,z}(\psi_0, \pi) D_\psi p_{yd,z}(\psi_0, \pi)',
$$

where $\psi \equiv (\beta, \zeta)$, $\psi_0 \equiv (0, \zeta_0)$ and $D_\psi p_{yd,z}(\psi_0, \pi) \equiv \partial p_{yd,z}(\psi_0, \pi)/\partial \psi$. The second one is

$$
K(\pi; \gamma_0) = - \sum_{y,d,z=0,1} \frac{\phi_{z,0}}{p_{yd,z}(\theta_0)} \frac{\partial p_{yd,z}(\theta_0)}{\partial \beta_0} D_\psi p_{yd,z}(\psi_0, \pi).
$$

Finally, $G(\cdot; \gamma_0)$ is a mean zero Gaussian process indexed by $\pi \in \Pi = [-0.99, 0.99]$ with bounded continuous sample paths and covariance kernel for $\pi_1, \pi_2 \in \Pi$ equal to

$$
\Omega(\pi_1, \pi_2; \gamma_0) = S_\psi V^\dagger((\psi_0, \pi_1), (\psi_0, \pi_2); \gamma_0) S_\psi',
$$

where $S_\psi \equiv [I_{d_\psi} : 0_{d_\psi \times 1}]$ is a selector matrix that selects the subvector ψ from θ and

$$
\begin{aligned}
V^\dagger(\theta_1, \theta_2; \gamma_0) \equiv{}& E_{\gamma_0}\left[\left(\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i) \frac{D_\theta p_{yd,z}^\dagger(\theta_1)}{p_{yd,z}(\theta_1)}\right)\left(\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i) \frac{D_\theta p_{yd,z}^\dagger(\theta_2)'}{p_{yd,z}(\theta_2)}\right)\right] \\
&- E_{\gamma_0}\left[\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i) \frac{D_\theta p_{yd,z}^\dagger(\theta_1)}{p_{yd,z}(\theta_1)}\right] E_{\gamma_0}\left[\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i) \frac{D_\theta p_{yd,z}^\dagger(\theta_2)'}{p_{yd,z}(\theta_2)}\right]
\end{aligned}
$$

35

$$= \sum_{y,d,z=0,1} \frac{p_{yd,z}(\theta_0)\phi_{z,0}}{p_{yd,z}(\theta_1)p_{yd,z}(\theta_2)} D_\theta p_{yd,z}^\dagger(\theta_1) p_{yd,z}^\dagger(\theta_2)'$$

$$- \left(\sum_{y,d,z=0,1} \frac{p_{yd,z}(\theta_0)\phi_{z,0}}{p_{yd,z}(\theta_1)} D_\theta p_{yd,z}^\dagger(\theta_1) \right) \left(\sum_{y,d,z=0,1} \frac{p_{yd,z}(\theta_0)\phi_{z,0}}{p_{yd,z}(\theta_2)} D_\theta p_{yd,z}^\dagger(\theta_2)' \right)$$

with $D_\theta p_{yd,z}^\dagger(\theta) \equiv B^{-1}(\beta)\partial p_{yd,z}(\theta)/\partial\theta$.

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. Here we specialize the results to the model that uses the Ali-Mikhail-Haq copula defined in (3.10). Figures 1–4 provide the simulated finite-sample density functions of the estimators of the threshold-crossing model parameters in red and their asymptotic approximations in blue. For the finite-sample distributions, we examine the true parameter values $\beta \in \{0, 0.1, 0.2, 0.4\}$, $\boldsymbol{\zeta} = 0.2$ and $\boldsymbol{\pi} = (0.6, 0.4, 0.4)$. Under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ the asymptotic distributional approximations use the corresponding parameter values with $b = \sqrt{n}\beta$, $\zeta_0 = \zeta$ and $\boldsymbol{\pi}_0 = \boldsymbol{\pi}$. Since $\hat{\boldsymbol{\theta}}_n = (\hat{\beta}_n, \hat{\boldsymbol{\mu}}_n) = (\hat{\beta}_n, h(\hat{\zeta}_n, \hat{\pi}_n))$, we use the distributions of the elements of $\beta + \tau_{0,b}^\beta(\pi_{0,b}^*)/\sqrt{n}$ and $h(\zeta_0 + \tau_{0,b}^\zeta(\pi_{0,b}^*)/\sqrt{n}, \pi_{0,b}^*)$ as our asymptotic approximations to the finite sample distributions of the elements of $\hat{\beta}_n$ and $\hat{\boldsymbol{\mu}}_n$. This approximation is asymptotically equivalent to using the limiting objects in Corollary 4.1(i) but performs better in finite samples by capturing the additional "randomness" arising from the $\sqrt{n}$-consistently estimable parameter $\hat{\zeta}_n$ in the distribution of $\hat{\boldsymbol{\mu}}_n$. 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.

7.2 Asymptotic Distributional Approximations for Wald Statistics

Similarly to the previous subsection, we now describe the additional quantities needed to obtain the asymptotic distributions of the Wald statistics in the threshold-crossing model example. The derivations can similarly be found in Appendix B.

Recalling the function λ is defined in expression (5.20) of AC14, the only new object appearing in $\lambda(\pi_{0,b}^*; \gamma_0, b)$ in Proposition 5.1 that is not a function of the specific restrictions under test $r(\cdot)$ or objects described in the previous subsection is the deterministic function $\Sigma(\pi; \gamma_0)$. For the threshold-crossing model, this function is given by $\Sigma(\pi; \gamma_0) = V^{-1}(\psi_0, \pi; \gamma_0)$, where

$$V(\psi_0, \pi; \gamma_0) = \sum_{y,d,z=0,1} \frac{\phi_{z,0}}{p_{yd,z}(\theta_0)} D_\theta p_{yd,z}^\dagger(\psi_0, \pi) D_\theta p_{yd,z}^\dagger(\psi_0, \pi)'.$$

Similarly to the previous subsection, we provide a brief simulation study to illustrate how well the random variable $\lambda(\pi_{0,b}^*; \gamma_0, b)$ from Proposition 5.1, arising as the limit of the Wald statistic under weak identification, approximates its finite-sample counterparts. Figures 5–8 provide the simulated finite sample density functions of $W_n(v)$ for one-dimensional null hypotheses on the separate elements of the parameter vector $\boldsymbol{\theta}$. This type of null hypothesis is a special case of those satisfying Assumptions R1–R2 in Section 5. 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 $\lambda(\pi_{0,b}^*; \gamma_0, b)$ are given in blue. In addition, the solid black line graphs the density function of a χ_1^2 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_1^2 approximations in the presence of weak/non-identification; (ii) as β grows larger, the distributions become approximately χ_1^2; and (iii) the new asymptotic distributional approximation perform very well, especially compared to the usual χ_1^2 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 $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_3$), the corresponding distributions in Figure 7 look well-approximated by the χ_1^2 distribution. This is perhaps due to the self-normalizing nature of Wald statistics.

7.3 Power Performance for One-Dimensional Robust Wald Tests

In this subsection, we provide a brief analysis of the power of one of our proposed robust Wald tests when applied to the one-dimensional parameter $\boldsymbol{\pi}_2$ of the threshold crossing model. Since the current literature does not contain tests with proven uniform size control for directly testing one-dimensional hypotheses in the maximum likelihood setting, we can only compare the power of our robust Wald test to a projected version of a full vector test. And since this model is estimated by maximum likelihood, the only test we could find in the literature *for the full parameter vector* $\boldsymbol{\theta}$ with proven asymptotic size control is the singularity-robust Anderson Rubin (SR-AR) test of Andrews and Guggenberger (2014) that uses the score function of the log-likelihood as the moment function. Thus, as a baseline performance measure, we compare the power of our robust test to the projected version of the SR-AR test.[27]

---

[27]Specifically, we minimize the SR-AR statistic over the remaining nuisance parameters β, $\boldsymbol{\zeta}$, $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_3$ and compare it to $\chi_5^2(0.95)^{-1}$.

For testing the null hypothesis, $H_0 : \boldsymbol{\pi}_2 = 0.4$ at the $\alpha = 0.05$ level, we examine the power of the robust Wald test that uses the (modified and) adjusted-Bonferroni CV described in Section 6.2, where we implement the CV with the second $\hat{I}_n^a(\hat{b}_n, \hat{\pi}_n)$ set described there with $a = 0.5$. We examine power under both weak and strong identification, corresponding here to $\beta = 0.2$ and 0.4. For these two values of β, the finite sample distributions of the data are generated identically to those in Sections 7.2-7.3 except that in order to produce power curves, we vary the true underlying value of $\boldsymbol{\pi}_2$ across a space of alternative hypotheses. These power curves, along with those of the projected SR-AR test are shown in Figure 9. Here, we can see the clear dominance of the robust Wald test in comparison to projected SR-AR under strong identification. Under weak identification, though the robust Wald test does not dominate, it exhibits higher power over most of the alternative space, with especially pronounced power differences occurring at more local alternatives.

8 Empirical Application: The Effect of Education on Crime

We now provide a short identification-robust empirical analysis that revisits some of the analysis of Lochner and Moretti (2004) on how educational attainment affects an individual's subsequent participation in crime. For this application, we use US Census data (Lochner and Moretti's, 2004 "inmates" data). Of the many sets of variables examined by these authors, one fits particularly neatly into the threshold crossing model of a triangular system (Example 2.3) we examine in detail in this paper. In terms of the variables of this model, Y_i is an indicator variable that equals one if the individual is in prison (labeled "prison" in the authors' dataset), D_i is an indicator variable that equals one if the individual is a high school dropout (labeled "drop") and Z_i is an indicator variable that equals one if the individual's high school required at least 11 years of schooling (labeled "ca11"). All data and descriptions thereof are freely available on Enrico Moretti's website (http://eml.berkeley.edu// moretti/).

We focus on the subpopulation of black individuals. Lochner and Moretti (2004) also provide separate analyses for white vs black individuals. We further focus on the subpopulation of black individuals turning age 14 in 1958 or later to account both for the impact of the Supreme Court decision *Brown v. Board of Education* and to mitigate cohort and/or time effects (see Lochner and Moretti, 2004 for further details). This leaves us with a final subpopulation of $n = 184,171$ individuals.

From this subpopulation, the maximum likelihood point estimates of the threshold crossing model parameters are as follows: $\hat{\beta}_n = -0.0137$, $\hat{\zeta}_n = 0.3060$, $\hat{\boldsymbol{\pi}}_{1,n} = 0.0260$, $\hat{\boldsymbol{\pi}}_{2,n} = 0.0782$ and $\hat{\boldsymbol{\pi}}_{3,n} = 0.0394$. Loosely speaking, note that the value of $\hat{\beta}_n$ may be indicative of weak

identification since $|\sqrt{n}\hat{\beta}_n| = 5.879$, roughly in line with b values that produce nonstandard densities in our simulation analysis of Sections 7.1–7.2. We perform robust Wald inference for the parameter $\boldsymbol{\pi}_2$, the counterfactual probability that an individual would be incarcerated had they dropped out of high school. To perform inference, we use the same (modified and) adjusted-Bonferroni CV for $\alpha = 0.05$ as described in Section 7.3, yielding a CV $c_{1-\alpha,n}^{AB} \approx 11.5$.[28] Forming a robust confidence interval for $\boldsymbol{\pi}_2$, by finding all hypothesized values of $\boldsymbol{\pi}_2$ that are not rejected by the robust Wald test, we obtain a 95% confidence interval equal to $[lb^*, 0.326]$, where $lb^* > 0$ is some small number that provides the lower bound on the true parameter space for $\boldsymbol{\pi}_2$. It is interesting to note that this implies that we fail to reject *any* small value of the counterfactual probability.

---

[28]Due to the structure of the parameter space, the CV does not depend upon the null hypothesized value for $\boldsymbol{\pi}_2$.

A   Appendix A: Proofs of Main Results

Proof of Theorem 3.1: When $\beta = 0$,

$$\frac{\partial Q_n(\theta)}{\partial \pi'} = \frac{\partial \Psi_n}{\partial \mathbf{g}'} \frac{\partial \bar{\mathbf{g}}_n(\beta, h(\mu))}{\partial \pi'} = \frac{\partial \Psi_n}{\partial \mathbf{g}'} \frac{\partial \bar{\mathbf{g}}_n(\beta, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}'} \frac{\partial h(\mu)}{\partial \pi'} = 0_{1 \times d_\pi}$$

for all $\theta = (0, \mu) \in \Theta \equiv \{(\beta, \mu) \in \mathbb{R}^{d_\theta} : (\beta, h(\mu)) \in \Theta\}$. ∎

Proof of Theorem 3.2 First note that

$$\frac{\partial g_n^{(1)}(0, \mu^{(1)})}{\partial \pi_1^{(1)}} = \frac{\partial \bar{\mathbf{g}}_n(0, h^{(1)}(\mu^{(1)}))}{\partial \pi_1^{(1)}} = \frac{\partial \bar{\mathbf{g}}_n(0, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}'} \bigg|_{\boldsymbol{\mu} = h^{(1)}(\mu^{(1)})} \times \frac{\partial h^{(1)}(\mu^{(1)})}{\partial \pi_1^{(1)}} = 0$$

by Steps 1 and 2. By way of induction, for $1 \leq i-1 \leq d_\pi - 1$, assume that the first $i-1$ columns of $\partial g_n^{(i-1)}(0, \mu^{(i-1)})/\partial \pi^{(i-1)\prime}$ are equal to zero. Then by Step 8 of the algorithm,

$$\frac{\partial g_n^{(i)}(0, \mu^{(i)})}{\partial \pi^{(i)\prime}} = \frac{\partial g_n^{(i-1)}(0, h^{(i)}(\mu^{(i)}))}{\partial \pi^{(i)\prime}} = \frac{\partial g_n^{(i-1)}(0, \mu^{(i-1)})}{\partial \mu^{(i-1)\prime}} \bigg|_{\mu^{(i-1)} = h^{(i)}(\mu^{(i)})} \times \frac{\partial h^{(i)}(\mu^{(i)})}{\partial \pi^{(i)\prime}}$$

$$= \left[\frac{\partial g_n^{(i-1)}(0, \mu^{(i-1)})}{\partial \zeta^{(i-1)\prime}} \vdots \frac{\partial g_n^{(i-1)}(0, \mu^{(i-1)})}{\partial (\pi_1^{(i-1)}, \dots, \pi_{i-1}^{(i-1)})} \vdots \frac{\partial g_n^{(i-1)}(0, \mu^{(i-1)})}{\partial (\pi_i^{(i-1)}, \dots, \pi_{d_\pi}^{(i-1)})} \right] \bigg|_{\mu^{(i-1)} = h^{(i)}(\mu^{(i)})}$$

$$\times \left[\frac{\partial h^{(i)}(\mu^{(i)})}{\partial (\pi_1^{(i)}, \dots, \pi_{i-1}^{(i)})} \vdots \frac{\partial h^{(i)}(\mu^{(i)})}{\partial \pi_i^{(i)}} \vdots \frac{\partial h^{(i)}(\mu^{(i)})}{\partial (\pi_{i+1}^{(i)}, \dots, \pi_{d_\pi}^{(i)})} \right]$$

$$= \left[\frac{\partial g_n^{(i-1)}(0, \mu^{(i-1)})}{\partial \zeta^{(i-1)\prime}} \vdots 0_{d_{\mathbf{g}} \times (i-1)} \vdots \frac{\partial g_n^{(i-1)}(0, \mu^{(i-1)})}{\partial (\pi_i^{(i-1)}, \dots, \pi_{d_\pi}^{(i-1)})} \right] \bigg|_{\mu^{(i-1)} = h^{(i)}(\mu^{(i)})}$$

$$\times \left[\begin{array}{c} 0_{(d_{\boldsymbol{\mu}} - d_\pi) \times (i-1)} \\ C^{(i)}(\mu^{(i)}) \\ 0_{(d_\pi - i + 1) \times (i-1)} \end{array} \quad \vdots \quad \frac{\partial h^{(i)}(\mu^{(i)})}{\partial \pi_i^{(i)}} \quad \vdots \quad \frac{\partial h^{(i)}(\mu^{(i)})}{\partial (\pi_{i+1}^{(i)}, \dots, \pi_{d_\pi}^{(i)})} \right]$$

$$= \left[0_{d_{\mathbf{g}} \times (i-1)} \vdots \frac{\partial g_n^{(i-1)}(0, \mu^{(i-1)})}{\partial \mu^{(i-1)\prime}} \frac{\partial h^{(i)}(\mu^{(i)})}{\partial \pi_i^{(i)}} \vdots \frac{\partial g_n^{(i-1)}(0, \mu^{(i-1)})}{\partial \mu^{(i-1)\prime}} \frac{\partial h^{(i)}(\mu^{(i)})}{\partial (\pi_{i+1}^{(i)}, \dots, \pi_{d_\pi}^{(i)})} \right] \bigg|_{\mu^{(i-1)} = h^{(i)}(\mu^{(i)})}$$

$$= \left[0_{d_{\mathbf{g}} \times (i-1)} \vdots 0_{d_{\mathbf{g}} \times 1} \vdots \frac{\partial g_n^{(i-1)}(0, \mu^{(i-1)})}{\partial \mu^{(i-1)\prime}} \frac{\partial h^{(i)}(\mu^{(i)})}{\partial (\pi_{i+1}^{(i)}, \dots, \pi_{d_\pi}^{(i)})} \right] \bigg|_{\mu^{(i-1)} = h^{(i)}(\mu^{(i)})},$$

where the third equality results from the definition of $\mu^{(i)}$ in Step 6, the fourth equality follows from Step 7 and the final equality follows from Steps 5 and 6.

Hence, we have shown that for $1 \leq i \leq d_\pi$, the first i columns of $\partial g_n^{(i)}(0, \mu^{(i)})/\partial \pi^{(i)\prime}$ are

equal to zero. In particular, $\partial g_n^{(d_\pi)*}(0, \mu^{(d_\pi)})/\partial \pi^{(d_\pi)\prime} = 0_{d_g \times d_\pi}$. Also note that Step 8 defines θ as equal to $(\beta, \mu^{(d_\pi)})$ and

$$
\begin{aligned}
\bar{g}_n(\theta) = \bar{g}_n(\beta, h^{(1)} \circ \ldots \circ h^{(d_\pi)}(\mu^{(d_\pi)})) &= g_n^{(1)}(\beta, h^{(2)} \circ \ldots \circ h^{(d_\pi)}(\mu^{(d_\pi)})) \\
= g_n^2(\beta, h^{(3)} \circ \ldots \circ h^{(d_\pi)}(\mu^{(d_\pi)})) &= \ldots = g_n^{(d_\pi)}(\beta, \mu^{(d_\pi)}),
\end{aligned}
$$

where the first equality follows from the definition of h in Step 8, the second equality follows from the definition of $g_n^{(1)}(\theta^{(1)})$ in Step 4 and the final two equalities follow from the definition of $g_n^{(i)}(\theta^{(i)})$ in Step 8. Thus for $\beta = 0$, using the definition of $h(\cdot)$ in Step 8, we have

$$
\begin{aligned}
\begin{bmatrix} \cdots & 0_{1 \times d_\pi} \\ \vdots & \vdots \\ \cdots & 0_{1 \times d_\pi} \end{bmatrix} &= \frac{\partial g_n^{(d_\pi)}(\theta^{(d_\pi)})}{\partial \mu^{(d_\pi)\prime}} = \frac{\partial \bar{g}_n(\beta, h^{(1)} \circ \ldots \circ h^{(d_\pi)}(\mu^{(d_\pi)}))}{\partial \mu^{(d_\pi)\prime}} \\
&= \frac{\partial \bar{g}(\beta, h(\mu))}{\partial \mu'} = \frac{\partial \bar{g}(\theta)}{\partial \mu'}\bigg|_{\theta = (\beta, h(\mu))} \times \frac{\partial h(\mu)}{\partial \mu'}
\end{aligned}
$$

so that $h : \mathcal{M} \to \mathcal{M}$ satisfies Procedure 3.1 if it is one-to-one. This latter property holds because each $\partial h^{(i)}(\mu^{(i)})/\partial \mu^{(i)\prime}$ for $i = 1, \ldots, d_\pi$ has full rank by Steps 3 and 7 and $h = h^{(1)} \circ \ldots \circ h^{(d_\pi)}$ by Step 8. ∎

Proof of Proposition 3.1 First, when $\beta = 0$, under Assumption ID, there exists at least one column in $\partial \bar{g}_n(\theta)/\partial \mu'$ that is linearly dependent on the other columns, which implies that there exists a nonzero vector $m^{(1)}$ such that (3.3) holds. Thus, (3.4) is a well-defined system of ODE's with an initial condition that is determined by constants of integration. By the (global) Picard-Lindelöf Theorem (Picard, 1893; Lindelöf, 1894), since $m^{(1)}(\cdot)$ is Lipschitz continuous on compact $\mathcal{M}^{(1)}$, there exists a solution $h^{(1)}$ on $\mathcal{M}^{(1)}$ of (3.4). Since the choice of constants of integration for this solution does not affect (3.4), it is always possible to choose them to ensure full rank of $\partial h^{(1)}(\mu^{(1)})/\partial \mu^{(1)\prime}$. Now by way of induction, for $1 \le i - 1 \le d_\pi - 1$, since $\partial h^{(i)}(\mu^{(i)})/\partial \mu^{(i)\prime}$ is full rank and $rank(\partial g_n^{(i-1)}(\theta^{(i-1)})/\partial \mu^{(i-1)\prime}) = r$, it follows that

$$
rank\left(\frac{\partial g_n^{(i)}(\theta^{(i)})}{\partial \mu^{(i)\prime}}\right) = rank\left(\frac{\partial g_n^{(i-1)}(\theta^{(i-1)})}{\partial \mu^{(i-1)\prime}} \frac{\partial h^{(i)}(\mu^{(i)})}{\partial \mu^{(i)\prime}}\right) = r.
$$

Thus, there exists a nonzero vector $m^{(i)}$ such that (3.5) holds. Given (3.6), since $m^{(i)}(\cdot)$ is Lipschitz continuous on compact $\mathcal{M}^{(i)}$, there exists a solution $h^{(i)}$ on $\mathcal{M}^{(i)}$. Similarly to before, since the choice of constants of integration for this solution does not affect (3.6), it is always possible to choose them to ensure (1) and (2) of Step 7 hold. Therefore, $h = h^{(1)} \circ \cdots \circ h^{(d_\pi)}$

exists on $\mathcal{M} = \mathcal{M}^{(d_\pi)}$. ∎

Proof of Lemma 4.1: For any $\mu \in \mathcal{M}$, since $M(h(\mu))$ has full rank, $\partial h(\mu)/\partial\mu'$ has full rank by Step 2 of Procedure 3.1. Therefore

$$\frac{\partial\bar{h}(\theta)}{\partial\theta'} = \left[\begin{array}{cc} 1 & 0 \\ 0 & \frac{\partial h(\mu)}{\partial\mu} \end{array} \right]$$

has full rank for any $\theta \in \Theta$. Also, since $h : \mathcal{M} \to \boldsymbol{\mathcal{M}}$ is proper, $\bar{h} : \Theta \to \boldsymbol{\Theta}$ is also proper. Combining these results with Assumption H(ii), we can apply Hadamard's global inverse function theorem Hadamard (1906a,b) to $\bar{h} : \Theta \to \boldsymbol{\Theta}$, and conclude that $\bar{h}$ is a homeomorphism. ∎

Proof of Lemma 4.2: Suppose Assumption Reg3*(v) holds. Without loss of generality we may permute the elements of μ^s so that

$$h_\pi^s(\mu) = \left(\begin{array}{cc} 0_{(d_s - \tilde{d}_\pi^*) \times \tilde{d}_\pi^*} & 0_{(d_s - \tilde{d}_\pi^*) \times (d_\pi - \tilde{d}_\pi^*)} \\ D(\mu) & 0_{\tilde{d}_\pi^* \times (d_\pi - \tilde{d}_\pi^*)} \end{array} \right),$$

where $D(\mu)$ is a diagonal full rank $\tilde{d}_\pi^* \times \tilde{d}_\pi^*$ matrix. By definition, the column space of $h_\pi^s(\mu)$ is equal to

$$\{v : v = h_\pi^s(\mu)x \text{ for some } x \in \mathbb{R}^{d_\pi}\}$$
$$= \{(0_{1 \times (d_s - \tilde{d}_\pi^*)}, v_2')' : v_2 \in \mathbb{R}^{\tilde{d}_\pi^*} \text{ and for each } i = 1, \ldots, \tilde{d}_\pi^*, \; v_{2,i} = D_{ii}(\mu)x_i \text{ for some } x_i \in \mathbb{R}\}$$
$$= \{(0_{1 \times (d_s - \tilde{d}_\pi^*)}, x_2)' : x_2 \in \mathbb{R}^{\tilde{d}_\pi^*},$$

which clearly satisfies the condition in Assumption Reg3*(iii) since it does not depend upon μ. ∎

The proofs of Theorem 4.1, Corollary 4.1 and Proposition 5.1 make use of the following auxiliary lemmas. The following lemma applies some of the main results of AC12.

Lemma A.1. *(i) Suppose Assumptions CF, Reg1 and ID, 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 $\|b\| < \infty$,*

$$\left(\begin{array}{c} \sqrt{n}(\hat{\beta}_n - \beta_n) \\ \sqrt{n}(\hat{\zeta}_n - \zeta_n) \\ \hat{\pi}_n \end{array} \right) \xrightarrow{d} \left(\begin{array}{c} \tau_{0,b}^\beta(\pi_{0,b}^*) \\ \tau_{0,b}^\zeta(\pi_{0,b}^*) \\ \pi_{0,b}^* \end{array} \right).$$

(ii) Suppose Assumptions CF, Reg1 and ID, 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 $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

$$\sqrt{n} \begin{pmatrix} \hat{\beta}_n - \beta_n \\ \hat{\zeta}_n - \zeta_n \\ \iota(\beta_n)(\hat{\pi}_n - \pi_n) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z_\beta \\ Z_\zeta \\ Z_\pi \end{pmatrix}.$$

Proof. Theorem 3.1 directly implies that Assumption A of AC12 holds when applied to the θ and $Q_n(\theta)$ of this paper. Then (i) and (ii) follow by direct application of Theorems 3.1(a) and 3.2(a) of AC12. ∎

The next lemma ensures we can write $\hat{\boldsymbol{\theta}}_n = (\hat{\beta}_n, h(\hat{\mu}_n))$.

Lemma A.2. *Suppose Assumption H holds. Then, $\hat{\boldsymbol{\theta}}_n = (\hat{\beta}_n, h(\hat{\mu}_n))$ for some $\hat{\theta}_n = (\hat{\beta}_n, \hat{\mu}_n) \in \Theta$ such that $Q_n(\hat{\theta}_n) = \inf_{\theta \in \Theta} Q_n(\theta) + o(n^{-1})$.*

Proof. The reparameterization function $\bar{h} : \Theta \to \boldsymbol{\Theta}$ is bijective by Lemma 4.1, which implies $\boldsymbol{\Theta} = \bar{h}(\Theta)$ and $\Theta = h^{-1}(\boldsymbol{\Theta})$ so that

$$\boldsymbol{Q}_n(\hat{\boldsymbol{\theta}}_n) = \inf_{\boldsymbol{\theta} \in \bar{h}(\Theta)} \boldsymbol{Q}_n(\boldsymbol{\theta}) + o(n^{-1}) = \inf_{\bar{h}^{-1}(\boldsymbol{\theta}) \in \Theta} \boldsymbol{Q}_n(\bar{h}(\bar{h}^{-1}(\boldsymbol{\theta}))) + o(n^{-1})$$

$$= \inf_{\bar{h}^{-1}(\boldsymbol{\theta}) \in \Theta} Q_n(\bar{h}^{-1}(\boldsymbol{\theta})) + o(n^{-1})$$

$$= \inf_{\theta \in \Theta} Q_n(\theta) + o(n^{-1}) = Q_n(\hat{\theta}_n)$$

for some $\hat{\theta}_n \in \Theta$. ∎

Proof of Theorem 4.1: (i) Using Lemma A.2, begin by decomposing $\hat{\boldsymbol{\mu}}_n^s - \boldsymbol{\mu}_n^s = h^s(\hat{\mu}_n) - h^s(\mu_n)$ as follows:

$$h^s(\hat{\mu}_n) - h^s(\mu_n) = [h^s(\hat{\zeta}_n, \hat{\pi}_n) - h^s(\zeta_n, \hat{\pi}_n)] + [h^s(\zeta_n, \hat{\pi}_n) - h^s(\zeta_n, \pi_n)]$$

$$= h^s_\zeta(\hat{\mu}_n)(\hat{\zeta}_n - \zeta_n) + [h^s(\zeta_n, \hat{\pi}_n) - h^s(\zeta_n, \pi_n)] + o_p(n^{-1/2}),$$

where the second equality uses a mean value expansion (with respect to ζ) that holds by Lemma A.1(i) and Lemma 4.1(ii). Using this decomposition, we have

$$\begin{pmatrix} \sqrt{n}(\hat{\beta}_n - \beta_n) \\ \sqrt{n}\tilde{A}_1(\hat{\mu}_n)(\hat{\boldsymbol{\mu}}_n^s - \boldsymbol{\mu}_n^s) \\ \tilde{A}_2(\hat{\mu}_n)(\hat{\boldsymbol{\mu}}_n^s - \boldsymbol{\mu}_n^s) \end{pmatrix} = \begin{pmatrix} \sqrt{n}(\hat{\beta}_n - \beta_n) \\ \sqrt{n}\tilde{A}_1(\hat{\mu}_n)h^s_\zeta(\hat{\mu}_n)(\hat{\zeta}_n - \zeta_n) \\ \tilde{A}_2(\hat{\mu}_n)[h^s(\zeta_n, \hat{\pi}_n) - h^s(\zeta_n, \pi_n)] \end{pmatrix}$$

43

$$+ \begin{pmatrix} 0 \\ \sqrt{n}\tilde{A}_1(\hat{\mu}_n)[h^s(\zeta_n, \hat{\pi}_n) - h^s(\zeta_n, \pi_n)] \\ \tilde{A}_2(\hat{\mu}_n)h^s_\zeta(\hat{\mu}_n)(\hat{\zeta}_n - \zeta_n) \end{pmatrix} + o_p(1)$$

$$= \begin{pmatrix} \sqrt{n}(\hat{\beta}_n - \beta_n) \\ \tilde{A}_1(\hat{\mu}_n)h^s_\zeta(\hat{\mu}_n)\sqrt{n}(\hat{\zeta}_n - \zeta_n) + \tilde{\eta}^*_{0,b} \\ \tilde{A}_2(\hat{\mu}_n)[h^s(\zeta_n, \hat{\pi}_n) - h^s(\zeta_n, \pi_n)] \end{pmatrix} + o_p(1)$$

$$\xrightarrow{d} \begin{pmatrix} \tau^\beta_{0,b}(\pi^*_{0,b}) \\ \tilde{A}_1(\zeta_0, \pi^*_{0,b})h^s_\zeta(\zeta_0, \pi^*_{0,b})\tau^\zeta_{0,b}(\pi^*_{0,b}) + \tilde{\eta}^*_{0,b} \\ \tilde{A}_2(\zeta_0, \pi^*_{0,b})[h^s(\zeta_0, \pi^*_{0,b}) - \boldsymbol{\mu}^s_0] \end{pmatrix}$$

under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$, where the second equality follows from Assumptions Reg2 and Reg3, Lemma A.1(i) and the CMT and the weak convergence follows from Assumption Reg2, Lemma A.1(i), the CMT and the fact that $h^s(\zeta_0, \pi_0) = \boldsymbol{\mu}^s_0$.

(ii) For the $\beta_0 = 0$ case, the same decomposition of $\hat{\boldsymbol{\mu}}^s_n - \boldsymbol{\mu}^s_n = h^s(\hat{\mu}_n) - h^s(\mu_n)$ as that used in the proof of part (i) and similar reasoning imply

$$\sqrt{n} \begin{pmatrix} \hat{\beta}_n - \beta_n \\ \tilde{A}_1(\hat{\mu}_n)(\hat{\boldsymbol{\mu}}^s_n - \boldsymbol{\mu}^s_n) \\ \iota(\beta_n)\tilde{A}_2(\hat{\mu}_n)(\hat{\boldsymbol{\mu}}^s_n - \boldsymbol{\mu}^s_n) \end{pmatrix} = \begin{pmatrix} \sqrt{n}(\hat{\beta}_n - \beta_n) \\ \tilde{A}_1(\hat{\mu}_n)h^s_\zeta(\hat{\mu}_n)\sqrt{n}(\hat{\zeta}_n - \zeta_n) \\ \tilde{A}_2(\hat{\mu}_n)\sqrt{n}\iota(\beta_n)[h^s(\zeta_n, \hat{\pi}_n) - h^s(\zeta_n, \pi_n)] \end{pmatrix} + o_p(1).$$

A mean-value expansion, Lemma 4.1(ii) and the consistency of $\hat{\mu}_n$ under $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ given by Lemma A.1(ii) provide that

$$\tilde{A}_2(\hat{\mu}_n)\sqrt{n}\iota(\beta_n)[h^s(\zeta_n, \hat{\pi}_n) - h^s(\zeta_n, \pi_n)] = \tilde{A}_2(\hat{\mu}_n)\sqrt{n}\iota(\beta_n)[(h^s_\pi(\zeta_n, \hat{\pi}_n) + o_p(1))(\hat{\pi}_n - \pi_n)]$$
$$= \tilde{A}_2(\hat{\mu}_n)h^s_\pi(\zeta_n, \hat{\pi}_n)\sqrt{n}\iota(\beta_n)(\hat{\pi}_n - \pi_n) + o_p(1),$$

where the second equality follows from Lemma 4.1(ii) and Lemma A.1(ii). Putting these results together, we have

$$\sqrt{n} \begin{pmatrix} \hat{\beta}_n - \beta_n \\ \tilde{A}_1(\hat{\mu}_n)(\hat{\boldsymbol{\mu}}^s_n - \boldsymbol{\mu}^s_n) \\ \iota(\beta_n)\tilde{A}_2(\hat{\mu}_n)(\hat{\boldsymbol{\mu}}^s_n - \boldsymbol{\mu}^s_n) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z_\beta \\ \tilde{A}_1(\mu_0)h^s_\zeta(\mu_0)Z_\zeta \\ \tilde{A}_2(\mu_0)h^s_\pi(\mu_0)Z_\pi \end{pmatrix}$$

by Assumption Reg2, Lemma A.1(ii) and the CMT. Finally, for the $\beta_0 \neq 0$ case, note that a standard mean value expansion for $\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}_n = h(\hat{\mu}_n) - h(\mu_n)$, Lemma 4.1(ii), Lemma A.1(ii)

and the CMT imply

$$\sqrt{n} \begin{pmatrix} \hat{\beta}_n - \beta_n \\ \hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}_n^s \end{pmatrix} = \begin{pmatrix} \sqrt{n}(\hat{\beta}_n - \beta_n) \\ \sqrt{n}h_\mu(\hat{\mu}_n)(\hat{\mu}_n - \mu_n) \end{pmatrix} + o_p(1)$$

$$= \begin{pmatrix} \sqrt{n}(\hat{\beta}_n - \beta_n) \\ h_\zeta(\hat{\mu}_n)\sqrt{n}(\hat{\zeta}_n - \zeta_n) + h_\pi(\hat{\mu}_n)\sqrt{n}(\hat{\pi}_n - \pi_n) \end{pmatrix} + o_p(1)$$

$$\xrightarrow{d} \begin{pmatrix} Z_\beta \\ h_\zeta(\mu_0)Z_\zeta + \iota(\beta_0)^{-1}h_\pi(\mu_0)Z_\pi \end{pmatrix}. \quad \blacksquare$$

Proof of Corollary 4.1: For case (i),

$$\sqrt{n}(\hat{\boldsymbol{\mu}}_n^1 - \boldsymbol{\mu}_n^1) = \sqrt{n}[h^1(\hat{\zeta}_n) - h^1(\zeta_n)] = h_\zeta^1(\hat{\zeta}_n)\sqrt{n}(\hat{\zeta}_n - \zeta_n) + o_p(1) \xrightarrow{d} h_\zeta^1(\zeta_0)\tau_{0,b}^\zeta(\pi_{0,b}^*),$$

where the first equality follows from Lemma A.2, the second equality follows from the mean value theorem, Lemma 4.1(ii) and Lemma A.1(i) and the weak convergence follows from the CMT, Lemma 4.1(ii) and Lemma A.1(i). The results for $\hat{\beta}_n$, $\hat{\boldsymbol{\mu}}_n^2$ and the joint convergence of the three components follow directly from Lemmas A.2 and A.1(i), Lemma 4.1(ii) and the CMT.

For case (ii), note that

$$\sqrt{n}\iota(\beta_n)(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}_n) = \sqrt{n}\iota(\beta_n)[h(\hat{\zeta}_n, \hat{\pi}_n) - h(\zeta_n, \pi_n)]$$

$$= \sqrt{n}\iota(\beta_n)[h(\hat{\zeta}_n, \hat{\pi}_n) - h(\zeta_n, \hat{\pi}_n)] + \sqrt{n}\iota(\beta_n)[h(\zeta_n, \hat{\pi}_n) - h(\zeta_n, \pi_n)]$$

$$= \sqrt{n}\iota(\beta_n)[h_\zeta(\hat{\mu}_n)(\hat{\zeta}_n - \zeta_n) + o_p(n^{-1/2})]$$

$$+ \sqrt{n}\iota(\beta_n)[h_\pi(\zeta_n, \hat{\pi}_n)(\hat{\pi}_n - \pi_n) + o_p(n^{-1/2}\iota(\beta_n)^{-1})]$$

$$= h_\pi(\zeta_n, \hat{\pi}_n)\sqrt{n}\iota(\beta_n)(\hat{\pi}_n - \pi_n) + o_p(1) \xrightarrow{d} h_\pi(\mu_0)Z_\pi,$$

where the first equality follows from Lemma A.2, the third equality follows from the mean value theorem, Lemma 4.1(ii) and Lemma A.1(ii), while the final equality and weak convergence result follow from the CMT, Lemma 4.1(ii) and Lemma A.1(ii). Nearly identical arguments to those used for case (i) provide that $\sqrt{n}(\hat{\boldsymbol{\mu}}_n^1 - \boldsymbol{\mu}_n^1) \xrightarrow{d} h_\zeta^1(\zeta_0)Z_\zeta$. Joint convergence of the three components immediately follows from Lemma A.1(ii). $\blacksquare$

Proof of Proposition 6.1: The proof is nearly identical to the proof of Theorem 5.1(b)(iv) of AC12, using Proposition 5.1 in the place of Theorems 4.2 and 4.3 of AC12. $\blacksquare$

Proof of Proposition 6.2: The proof of this proposition verifies that the assumptions of Theorem Bonf-Adj of McCloskey (2017) hold, with some modifications. First, Assumption PS of McCloskey (2017) holds with $\gamma_1 = (\beta, \pi)$, $\gamma_2 = (\zeta, \delta)$ and $\gamma_3 = \varphi$. 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 (2017), 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 (2017), $T_n(\theta_n) = W_n(v_n)$ $\hat{h}_{n,1} = (\hat{b}_n, \hat{\pi}_n)$ and $\hat{h}_{n,2} = (\hat{\zeta}_n, \hat{\delta}_n)$. Proposition 5.1 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 (2017), $W_h = \lambda(\pi_{0,b}^*; \gamma_0, b)$ when $\|h_{1,1}\| < \infty$ and W_h is distributed $\chi^2_{d_r}$ when $\|h_{1,1}\| = \infty$. Lemma A.1 and Assumption FD provide the marginal weak convergence of $\hat{h}_{\omega_n} = (\hat{h}_{\omega_n,1}, \hat{h}_{\omega_n,2})$ for all sequences $\{\gamma_{\omega_n,h}\}$, where in the notation of McCloskey (2017), $\tilde{h}_1 = (b + \tau_{0,b}^\beta(\pi_{0,b}^*), \pi_{0,b}^*)$ when $\|h_{1,1}\| < \infty$, $\tilde{h}_1 = (b + Z_\beta, \pi_0)$ when $\|h_{1,1}\| = \infty$ and $h_2 = (\zeta_0, \delta_0)$. Joint convergence of $(T_{\omega_n}(\theta_{\omega_n}), \hat{h}_{\omega_n})$ follows from nearly identical arguments for joint convergence to those used in the proof of Theorem 5.1 of AC14.

Third, for Definition MLLD of McCloskey (2017), we are in what McCloskey (2017) refers to as "the usual case" for which $u = 1$, $\tilde{W}_h^{(1)} = \lambda(\pi_{0,b}^*; \gamma_0, b)$ and $\bar{H}^{(1),c} = \emptyset$ since $P(|\lambda(\pi_{0,b}^*; \gamma_0, b)| < \infty) = 1$ under the assumptions of Proposition 5.1. 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 $W_h = \lambda(\pi_{0,b}^*; \gamma_0, b) = \tilde{W}_h^{(1)}$ when $\|h_{1,1}\| < \infty$ and $W_h = \tilde{W}_h^{(1)}$ is distributed $\chi^2_{d_r}$ when $\|h_{1,1}\| = \infty$, the only item left to verify is that $\lambda(\pi_{0,b}^*; \gamma_0, b)$ 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 (2017), $\bar{H}^{(1)} = H$. This assumption holds for any $\underline{\delta}^{(1)} > 0$ and $\bar{\delta}^{(1)} \leq \alpha$ since $\lambda(\pi_{0,b}^*; \gamma_0, b)$ is an absolutely continuous random variable with quantiles that are continuous in b and π_0 and $\lambda(\pi_{0,b}^*; \gamma_0, b) \overset{d}{\sim} \chi^2_{d_r}$ for any b such that $\|b\| = \infty$. Fifth, Assumption Sel holds trivially since we are in the "usual case".

Sixth, Assumption CS of McCloskey (2017) can be modified and applied to $\hat{I}_n^a(\cdot)$ and its limit counterpart $I_0^a(\cdot)$ so that: (i)

$$\sup_{(b,\pi_0) \in \{(\tilde{b},\tilde{\pi}) \in \mathbb{R}^{d_\beta + d_\pi}_\infty : (\tilde{b},\tilde{\gamma}) \in \mathcal{L}\}} d_H(\hat{I}_n^a(b, \pi_0), I_0^a(b, \pi_0)) \overset{p}{\longrightarrow} 0$$

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) $I_0^a(\cdot)$ is a continuous and compact-valued correspondence; (iii) $P_{\gamma_n}(\hat{I}_n^a(\hat{b}_n, \hat{\pi}_n) \subset \bar{H}_1^{(1)}(\hat{h}_{n,2}^c)) = 1$ for all $n \geq 1$ and $\{\gamma_n\} \in \Gamma(\gamma_0)$ and $P(I_0^a(b + \tau_{0,b}^\beta(\pi_{0,b}^*), \pi_{0,b}^*) \subset \bar{H}_1^{(1)}(h_2^c)) = 1$; and (iv) $I_0^a(b + \tau_{0,b}^\beta(\pi_{0,b}^*), \pi_{0,b}^*)$ need not satisfy a coverage requirement (i.e., $P(h_1 \in I_0^a(b + \tau_{0,b}^\beta(\pi_{0,b}^*), \pi_{0,b}^*) \geq 1 - a)$. The proof of Theorem Bonf-Adj in McCloskey (2017) still goes through with this modification of Assumption CS. Condition (i) is satisfied by the consistency of $(\hat{\zeta}_n, \hat{\delta}_n)$

and the uniform consistency of $\hat{\Sigma}_n(\cdot)$ under any $\{\gamma_n\} \in \Gamma(\gamma_0)$. The former holds by Lemma A.1 and Assumption FD while the latter holds by Assumptions V1 and V2 of AC12. For condition (ii), $I_0^a(\cdot)$ is clearly continuous and compact-valued. Note that $\mathcal{P}(\hat{\zeta}_n, \hat{\delta}_n)$ and $\mathcal{P}(\zeta_0, \delta_0)$ are equal to $\bar{H}^{(1)}(\hat{h}_{n,2}^c)$ and $\bar{H}^{(1)}(h_2^c)$ in the notation of McCloskey (2017) 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, 2017), we are fixing the quantile at level $1 - \alpha$ and adding a size-correction function $\varsigma(\cdot)$ to it. The proof of Theorem Bonf-Adj of McCloskey (2017) can be easily adjusted to this modification. Rather than requiring the quantile adjustment function to be continuous, the proof requires $\varsigma(\cdot)$ to be continuous. That is, Assumption a(i) of McCloskey (2017) may be replaced by the analogous assumption: $\varsigma(\cdot)$ is continuous. In practice, $\varsigma(\cdot)$ is only evaluated at the point $(\hat{\zeta}_n, \hat{\delta}_n, \hat{\bar{\Sigma}}_n)$, which is consistent with this assumption. Due to the replacement of quantile adjustment by additive size-correction, Assumption a(ii) of McCloskey (2017) should also be replaced by the analogous assumption: $P(\lambda(\pi_{0,b}^*; \gamma_0, b) \geq \sup_{\ell \in \mathcal{L}_0^a(b,\gamma_0) \cap \mathcal{L}(v)} c_{1-\alpha}(\ell) + \varsigma(\zeta_0, \delta_0, \bar{\Sigma}(b, \gamma_0))) \leq \alpha$ for all $(b, \gamma_0) \in \mathcal{L}_0 \cap \mathcal{L}(v)$. This assumption holds by the construction of $\varsigma(\hat{\zeta}_n, \hat{\delta}_n, \hat{\bar{\Sigma}}_n)$ and the (uniform) consistency of $(\hat{\zeta}_n, \hat{\delta}_n, \hat{\Sigma}_n(\cdot))$.

Finally, Assumption Inf-Adj of McCloskey (2017) holds vacuously since $\bar{H}^{(1),c} = \emptyset$ and Assumption LB-Adj of that paper is imposed by Assumption DF2. ∎

B   Appendix B: Assumption Verifications for Threshold-Crossing Example

Before proceeding to verify the assumptions imposed for the Threshold-Crossing Model example, we provide the details for the claim that $\|\tilde{\eta}(\hat{\mu}_n)\|$ diverges for $\hat{\boldsymbol{\mu}}_n^s = (\hat{\boldsymbol{\mu}}_{n,3}, \hat{\boldsymbol{\mu}}_{n,4})$ made in the continuation of Example 2.3 in Section 4.

Proof $\|\tilde{\eta}(\hat{\mu}_n)\|$ diverges in Example 2.3: Note that

$$
\begin{aligned}
\tilde{\eta}_n(\hat{\mu}_n) &= \sqrt{n}\mathcal{S}(\hat{\mu}_n)\left[h_3(\zeta_n, \hat{\pi}_n) - h_3(\zeta_n, \pi_n) + \frac{C_3(h_3(\hat{\mu}_n), \hat{\zeta}_{1,n}; \hat{\pi}_n)}{C_1(h_3(\hat{\mu}_n), \hat{\zeta}_{1,n}; \hat{\pi}_n)}(\hat{\pi}_n - \pi_n)\right] \\
&= \sqrt{n}\mathcal{S}(\hat{\mu}_n)\left[\frac{\zeta_{3,n}(\zeta_{1,n} - 1)(\zeta_{1,n} - \zeta_{3,n})}{(\zeta_{1,n} - \zeta_{3,n}\hat{\pi}_n + \zeta_{1,n}\zeta_{3,n}\hat{\pi}_n)(\zeta_{1,n} - \zeta_{3,n}\pi_n + \zeta_{1,n}\zeta_{3,n}\pi_n)}\right. \\
&\qquad\qquad\qquad \left. + \frac{C_3(h_3(\hat{\mu}_n), \hat{\zeta}_{1,n}; \hat{\pi}_n)}{C_1(h_3(\hat{\mu}_n), \hat{\zeta}_{1,n}; \hat{\pi}_n)}\right](\hat{\pi}_n - \pi_n) \\
&= \sqrt{n}\mathcal{S}(\hat{\mu}_n)\left[\frac{\zeta_{3,n}(\zeta_{1,n} - 1)(\zeta_{1,n} - \zeta_{3,n})}{(\zeta_{1,n} - \zeta_{3,n}\hat{\pi}_n + \zeta_{1,n}\zeta_{3,n}\hat{\pi}_n)(\zeta_{1,n} - \zeta_{3,n}\pi_n + \zeta_{1,n}\zeta_{3,n}\pi_n)}\right. \\
&\qquad\qquad\qquad \left. - \frac{\hat{\zeta}_{3,n}(\hat{\zeta}_{1,n} - 1)(\hat{\zeta}_{1,n} - \hat{\zeta}_{3,n})}{(\hat{\zeta}_{1,n} - \hat{\zeta}_{3,n}\hat{\pi}_n + \hat{\zeta}_{1,n}\hat{\zeta}_{3,n}\hat{\pi}_n)^2}\right](\hat{\pi}_n - \pi_n) \\
&= \sqrt{n}\mathcal{S}(\hat{\mu}_n)\left[\frac{\tilde{\eta}_n^N(\hat{\mu}_n)}{\tilde{\eta}_n^D(\hat{\mu}_n)}\right](\hat{\pi}_n - \pi_n),
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{\eta}_n^N(\hat{\mu}_n) &= \zeta_{3,n}(\zeta_{1,n} - 1)(\zeta_{1,n} - \zeta_{3,n})(\hat{\zeta}_{1,n} - \hat{\zeta}_{3,n}\hat{\pi}_n + \hat{\zeta}_{1,n}\hat{\zeta}_{3,n}\hat{\pi}_n) \\
&\quad - \hat{\zeta}_{3,n}(\hat{\zeta}_{1,n} - 1)(\hat{\zeta}_{1,n} - \hat{\zeta}_{3,n})(\zeta_{1,n} - \zeta_{3,n}\pi_n + \zeta_{1,n}\zeta_{3,n}\pi_n) \\
&= \zeta_{3,n}(\zeta_{1,n} - 1)(\zeta_{1,n} - \zeta_{3,n})[(\hat{\zeta}_{1,n} - \hat{\zeta}_{3,n}\hat{\pi}_n + \hat{\zeta}_{1,n}\hat{\zeta}_{3,n}\hat{\pi}_n) - (\zeta_{1,n} - \zeta_{3,n}\pi_n + \zeta_{1,n}\zeta_{3,n}\pi_n)] \\
&\quad + [\zeta_{3,n}(\zeta_{1,n} - 1)(\zeta_{1,n} - \zeta_{3,n}) - \hat{\zeta}_{3,n}(\hat{\zeta}_{1,n} - 1)(\hat{\zeta}_{1,n} - \hat{\zeta}_{3,n})](\zeta_{1,n} - \zeta_{3,n}\pi_n + \zeta_{1,n}\zeta_{3,n}\pi_n) \\
&= \zeta_{3,n}^2(\zeta_{1,n} - 1)^2(\zeta_{1,n} - \zeta_{3,n})(\hat{\pi}_n - \pi_n) + O_p(n^{-1/2}) = O_p(n^{-1/2}\|\beta_n\|^{-1})
\end{aligned}
$$

with the final two equalities resulting from Lemma A.1 and a mean value expansion of the term $\hat{\zeta}_{3,n}(\hat{\zeta}_{1,n} - 1)(\hat{\zeta}_{1,n} - \hat{\zeta}_{3,n})$, and

$$
\tilde{\eta}_n^D(\hat{\mu}_n) = (\zeta_{1,n} - \zeta_{3,n}\hat{\pi}_n + \zeta_{1,n}\zeta_{3,n}\hat{\pi}_n + O_p(n^{-1/2}))^2(\zeta_{1,n} - \zeta_{3,n}\pi + \zeta_{1,n}\zeta_{3,n}\pi) = O_p(1)
$$

by Lemma A.1. Noting that both $\mathcal{S}(\hat{\mu}_n)$ and $\tilde{\eta}_n^D(\hat{\mu}_n)^{-1}$ are also $O_p(1)$ by Lemma A.1, we may combine the expressions for $\tilde{\eta}_n(\hat{\mu}_n)$, $\mathcal{S}(\hat{\mu}_n)$, $\tilde{\eta}_n^N(\hat{\mu}_n)$ and $\tilde{\eta}_n^D(\hat{\mu}_n)$ to conclude that $\|\tilde{\eta}_n(\hat{\mu}_n)\| =$

$\|O_p(n^{-1/2}\|\beta_n\|^{-1})\sqrt{n}(\hat{\pi}_n - \pi_n)\| = \|O_p(n^{-1/2}\|\beta_n\|^{-2})\| \to \infty$, according to Lemma A.1. ■

We now proceed to verify the imposed assumptions for the Threshold-Crossing Model example. Hereafter, Andrews and Cheng (2013a) and Han and Vytlacil (2017) are abbreviated as AC13 and HV16. The supplemental material for AC12, AC13 and AC14, Andrews and Cheng (2012b, 2013b, 2014b), are abbreviated as AC12supp, AC13supp and AC14supp. The working paper version of AC13 is abbreviated as ACMLwp. And "with respect to" is abbreviated as "w.r.t."

B.1 Assumptions for Threshold Crossing Models

The assumptions in the main text of the current paper and the assumptions in AC12 on objects involving the transformed parameter θ are verified under assumptions introduced in this section. The assumptions in AC12 are verified by verifying those in AC13.

Assumption TC1: $\{W_i : i \geq 1\}$ is an i.i.d. sequence.

Assumption TC2: (i) $Z \perp (\varepsilon, \nu)$;
(ii) F_ε and F_ν are known marginal distributions of ε and ν, respectively, that are strictly increasing and absolutely continuous with respect to the Lebesgue measure such that $E[\varepsilon] = E[\nu] = 0$ and $Var(\varepsilon) = Var(\nu) = 1$;
(iii) $(\varepsilon, \nu)' \sim F_{\varepsilon\nu}(\varepsilon, \nu) = C(F_\varepsilon(\varepsilon), F_\nu(v); \pi)$ where $C : (0,1)^2 \to (0,1)$ is a copula known up to a scalar parameter $\pi \in \Pi$ such that $C(u_1, u_2, ; \pi)$ is three-times differentiable in $(u_1, u_2, \pi) \in (0,1)^2 \times \Pi$;
(iv) The copula $C(u_1, u_2, ; \pi)$ satisfies

$$C(u_1|u_2; \pi) \prec_S C(u_1|u_2; \pi') \text{ for any } \pi < \pi', \tag{B.1}$$

where "$\prec_S$" is a stochastic ordering defined in HV16 (Definition 3.2);
(v) $(1, Z)$ does not lie in a proper linear subspace of $\mathbb{R}^2$;
(vi) $\boldsymbol{\Theta}^*$ is open and convex.

Given the form of h in (3.8) with $c_4(\zeta)$ set equal to zero, we write $\pi = \boldsymbol{\pi}_3$ in this assumption and below. The conditions in TC2 are sufficient for (global) identification of $\boldsymbol{\theta}$ when $\beta \neq 0$. The argument is similar to that in HV16, except that the condition for the parameter space TC2(vi) is stronger than that in HV16.

For the next assumption, define $\boldsymbol{\Theta}^*_\delta \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}^* : |\beta| < \delta\}$ for some $\delta > 0$.

Assumption TC3: (i) $\boldsymbol{\Theta} \equiv \boldsymbol{\Theta}_{-\pi} \times \Pi$, and $\boldsymbol{\Theta}_{-\pi}$ and Π are compact and simply connected;

49

(ii) $int(\boldsymbol{\Theta}) \supset \boldsymbol{\Theta}^*$;

(iii) For some $\delta > 0$, $\boldsymbol{\Theta} \supset \left\{ \beta \in \mathbb{R}^{d_\beta} : |\beta| < \delta \right\} \times \boldsymbol{\mathcal{Z}}^0 \times \Pi \supset \boldsymbol{\Theta}_\delta^*$ for some non-empty open set $\boldsymbol{\mathcal{Z}}^0 \subset \mathbb{R}^{d_{\boldsymbol\mu} - d_\pi}$ and Π.

(iv) $h^{-1}(\boldsymbol{\mathcal{Z}}^0 \times \Pi) = \mathcal{Z}^0 \times \Pi$ for some non-empty open set $\mathcal{Z}^0 \subset \mathbb{R}^{d_{\boldsymbol\mu} - d_\pi}$.

As is typical, Assumption TC3(i)-(ii) will be satisfied by a proper choice of the optimization parameter space. For concreteness, we define

$$\boldsymbol{\Theta}^* \equiv \{\boldsymbol{\theta} = (\beta, \boldsymbol{\zeta}, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \pi) \in [-0.98, 0.98] \times [0.01, 0.99] \times [0.01, 0.99] \times [0.01, 0.99] \times [-0.99, 0.99]$$
$$: 0.01 \leq \beta + \boldsymbol{\zeta} \leq 0.99\} \tag{B.2}$$

and

$$\boldsymbol{\Theta} \equiv \{\boldsymbol{\theta} = (\beta, \boldsymbol{\zeta}, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \pi) \in [-0.98 - \epsilon, 0.98 + \epsilon] \times [0.01 - \epsilon, 0.99 + \epsilon] \times [0.01 - \epsilon, 0.99 + \epsilon]$$
$$\times [0.01 - \epsilon, 0.99 + \epsilon] \times [-0.99 - \epsilon, 0.99 + \epsilon] : 0.01 - \epsilon \leq \beta + \boldsymbol{\zeta} \leq 0.99 + \epsilon\} \tag{B.3}$$

for some $\epsilon > 0$ so that TC3(i)-(ii) is clearly satisfied for small enough ϵ. Given the definition (B.2), TC4 below also holds if we define the parameter space $\boldsymbol{\Phi}^*(\boldsymbol{\theta})$ of $\phi \equiv \phi_1$ as

$$\boldsymbol{\Phi}^*(\boldsymbol{\theta}) = \boldsymbol{\Phi}^* \equiv [0.01, 0.99]. \tag{B.4}$$

TC3(iii) is satisfied by setting

$$\boldsymbol{\mathcal{Z}}^0 \equiv (0.01 - \delta, 0.99 + \delta)^3$$

for $\delta < \epsilon/2$. For TC3(iv), let $\tilde{h}^{-1}(\boldsymbol{\zeta}, \boldsymbol{\pi}) = (h_1^{-1}(\boldsymbol{\zeta}, \boldsymbol{\pi}), h_2^{-1}(\boldsymbol{\zeta}, \boldsymbol{\pi}), h_3^{-1}(\boldsymbol{\zeta}, \boldsymbol{\pi}))$, the first three elements of (3.12). Note that $h_4(\zeta, \pi) = \pi$ (i.e., $\boldsymbol{\pi}_3 = \pi$) and for any given $\pi \in \Pi$, $\tilde{h}^{-1}(\boldsymbol{\mathcal{Z}}^0, \pi)$ does not depend on π. Thus, we may set $\mathcal{Z}^0 = \tilde{h}^{-1}(\boldsymbol{\mathcal{Z}}^0, \pi)$ for any $\pi \in \Pi$, noting that $\mathcal{Z}^0$ must be a non-empty open set by the continuity of the first three elements of $h(\cdot)$. The latter follows from TC2(iii) and (3.8) after setting $c_1(\zeta) = \zeta_1$, $c_2(\zeta) = \zeta_2$ and $c_3(\zeta) = \zeta_3$.

Assumption TC4: (i) $\boldsymbol{\Gamma}$ is compact and $\boldsymbol{\Gamma} = \{\boldsymbol{\gamma} = (\boldsymbol{\theta}, \phi) : \boldsymbol{\theta} \in \boldsymbol{\Theta}^*, \phi \in \boldsymbol{\Phi}^*(\boldsymbol{\theta})\}$;

(ii) $\forall \delta > 0$, $\exists \boldsymbol{\gamma} = (\beta, \boldsymbol{\mu}, \phi) \in \boldsymbol{\Gamma}$ with $0 < |\beta| < \delta$;

(iii) $\forall \boldsymbol{\gamma} = (\beta, \boldsymbol{\mu}, \phi) \in \boldsymbol{\Gamma}$ with $0 < |\beta| < \delta$ for some $\delta > 0$, $\boldsymbol{\gamma}_a = (a\beta, \boldsymbol{\mu}, \phi) \in \boldsymbol{\Gamma}$ $\forall a \in [0, 1]$.

Assumption TC4(ii) guarantees that the true parameter space includes a region where weak identification occurs and TC4(iii) ensures that $\boldsymbol{\Gamma}$ is consistent with the existence of $K(\theta; \gamma)$, defined later.

Assumption TC5: (i) $C(u_1, u_2, ; \pi)$ is bounded away from zero over $(0, 1)^2 \times \Pi$;

(ii) $0 < \phi_1 \equiv \text{Pr}_\gamma[Z = 1] < 1 \ \forall \gamma \in \Gamma$.

Lemma B.1. *TC5 and TC2(iii) imply the following: for $(y, d, z) \in \{0, 1\}^3$, $\forall \boldsymbol{\gamma} = (\boldsymbol{\theta}, \phi) \in \boldsymbol{\Gamma}$, and $\forall \gamma = (\theta, \phi) \in \Gamma$,*

(i) the first, second, and third order derivatives of $\boldsymbol{p}_{yd,z}(\boldsymbol{\theta})$ are bounded over $\boldsymbol{\Theta}$;

(ii) $\boldsymbol{p}_{yd,z}(\boldsymbol{\theta})$ is bounded away from zero over $\boldsymbol{\Theta}$ and $0 < \phi_1 < 1$;

(iii) $\bar{h}(\theta)$ is three-times differentiable on Θ;

(iv) $p_{yd,z}(\theta) \equiv \boldsymbol{p}_{yd,z}(\bar{h}(\theta))$ is three-times differentiable on Θ and the first, second, and third order derivatives of $p_{yd,z}(\theta)$ are bounded over Θ;

(v) $p_{yd,z}(\theta)$ is bounded away from zero over Θ.

Proof of Lemma B.1: (i) holds by TC2(iii), the fact that the domain $\boldsymbol{\Theta}$ is compact by TC3(i), and the definitions of $\boldsymbol{p}_{yd,z}(\boldsymbol{\theta})$. (ii) immediately holds by TC5. For (iii), given (3.9), TC2(iii) and TC3(i) imply that $h(\mu)$ is three-times differentiable in μ and hence $\bar{h}(\theta) = (\beta, h(\mu))$ is three-times differentiable in θ. Next, (iv) holds by (i), (iii), and the chain rule, and (v) trivially holds by (ii).

B.2 Verification of Assumptions in the Main Text

Assumptions CF, ID, Jac, and Reg3 are verified in the main text. Assumption Reg1 is satisfied with $\bar{\boldsymbol{g}}_n(\boldsymbol{\theta}) = \hat{\xi}_n - \boldsymbol{g}(\boldsymbol{\theta})$, where each element $\boldsymbol{p}_{yd,z}(\boldsymbol{\theta})$ of the vector $\boldsymbol{g}(\boldsymbol{\theta})$ is continuously differentiable by TC2(iii). For Assumption H, H(i) holds since its sufficient conditions that Θ is bounded and h is continuous hold by S2(v), verified below, and by Proposition 3.1, respectively. H(ii) is also trivially satisfied by TC3(i). For Reg2, $rank(h_\pi^s(\mu)) = 1$ if $h^s(\pi)$ contains $h_2(\pi)$, $h_3(\pi)$ or $h_4(\pi)$ and $rank(h_\pi^s(\mu)) = 0$ otherwise, as can be seen from the form of h in (3.8) upon setting $c_1(\zeta) = \zeta_1$, $c_2(\zeta) = \zeta_2$, $c_3(\zeta) = \zeta_3$ and $c_4(\zeta) = 0$.

B.3 Verification of Assumptions in Andrews and Cheng (2013)

In this section, given our transformed parameter θ and associated transformed objects, we verify the regularity conditions for the asymptotic theory of the ML estimator $\hat{\theta}_n$ in AC13. Specifically, we show that Assumptions TC1–TC5 are sufficient for Assumptions S1–S4, B1, B2, C6, C7, V1, and V2 of AC13. Then, under Assumptions B1 and B2, Assumptions S1–S3 of AC13 imply Assumptions A, B3, C1–C4, C8, and D1–D3 of AC12; see Lemma 9.1 in ACMLwp. Maintaining the same labels of AC13, below we rewrite the assumptions of AC13 before verifying them. Note that in our stylized threshold crossing model, β is scalar. Therefore we do not consider Assumptions S3* and V1* of AC13 which apply to the vector β case.

Assumption S1: $\{W_i : i \geq 1\}$ is an i.i.d. sequence.[29]

Assumption S2: (i) For some function $\rho(w, \theta) \in \mathbb{R}$, $Q_n(\theta) = n^{-1} \sum_{i=1}^n \rho(W_i, \theta)$, where $\rho(w, \theta)$ is twice continuously differentiable in θ on an open set containing $\Theta^* \ \forall w \in \mathcal{W}$.
(ii) $\rho(w, \theta)$ does not depend on π when $\beta = 0 \ \forall w \in \mathcal{W}$.
(iii) $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$, $E_{\gamma_0} \rho(W_i, \psi, \pi)$ is uniquely minimized by $\psi_0 \ \forall \pi \in \Pi$.
(iv) $\forall \gamma_0 \in \Gamma$ with $\beta_0 \neq 0$, $E_{\gamma_0} \rho(W_i, \theta)$ is uniquely minimized by θ_0.
(v) $\Psi(\pi)$ is compact $\forall \pi \in \Pi$, and Π and Θ are compact.
(vi) $\forall \epsilon > 0$, $\exists \delta > 0$ such that $d_H(\Psi(\pi_1), \Psi(\pi_2)) < \epsilon \ \forall \pi_1, \pi_2 \in \Pi$ with $|\pi_1 - \pi_2| < \delta$, where $d_H(\cdot, \cdot)$ is the Hausdorff metric.

Verification of S2(i): By TC2(iii), $\boldsymbol{p}_{yd,z}(\boldsymbol{\theta})$ is twice continuously differentiable in $\boldsymbol{\theta}$. Then, since $p_{yd,z}(\theta) \equiv \boldsymbol{p}_{yd,z}(\bar{h}(\theta))$ is twice continuously differentiable by Lemma B.1, so is $\rho(w, \theta) = -\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(w) \log p_{yd,z}(\theta)$. ∎

Verification of S2(ii): It is easy to see from (2.5)–(2.6) that, when $\beta = 0$, $\boldsymbol{p}_{yd,0}(\boldsymbol{\theta}) = \boldsymbol{p}_{yd,1}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$ and (y, d), which implies that $\boldsymbol{p}_{yd,0}(\bar{h}(\theta)) = \boldsymbol{p}_{yd,1}(\bar{h}(\theta))$ for all θ. Therefore

$$
\begin{aligned}
p_{11,1}(\theta) &= p_{11,0}(\theta) = \zeta_3, \\
p_{10,1}(\theta) &= p_{10,0}(\theta) = \zeta_2, \\
p_{01,1}(\theta) &= p_{01,0}(\theta) = \zeta_1 - \zeta_3,
\end{aligned}
\tag{B.5}
$$

where the second equality in each equation is from (7.1)–(7.2). Therefore $p_{yd,z}(\theta)$ does not depend on π when $\beta = 0$, and hence $\rho(w, \theta) = -\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(w) \log p_{yd,z}(\theta)$ does not depend on π. ∎

Verification of S2(iii): When $\beta_0 = 0$, for $\psi \neq \psi_0$ and for a given π,

$$
\begin{aligned}
E_{\gamma_0} \rho(W_i, \psi, \pi) - E_{\gamma_0} \rho(W_i, \psi_0, \pi) &= -\sum_{y,d,z=0,1} p_{yd,z}(\psi_0, \pi_0) \phi_{z,0} \log \frac{p_{yd,z}(\psi, \pi)}{p_{yd,z}(\psi_0, \pi)} \\
&\geq -\log \sum_{y,d,z=0,1} p_{yd,z}(\psi_0, \pi_0) \phi_{z,0} \frac{p_{yd,z}(\psi, \pi)}{p_{yd,z}(\psi_0, \pi)} \\
&= -\log \sum_{y,d,z=0,1} p_{yd,z}(\psi, \pi) \phi_{z,0} \\
&= 0,
\end{aligned}
$$

where the last equality holds since $\sum_{y,d} p_{yd,1}(\theta) = \sum_{y,d} p_{yd,0}(\theta) = 1$ and $\phi_{0,0} = 1 - \phi_{1,0}$, and

---

[29]This is actually a sufficient condition for Assumption S1 of AC13.

the second-to-last equality holds since

$$p_{yd,z}(\psi_0, \pi_0) = p_{yd,z}(\psi_0, \pi) \equiv p_{yd}^0 \tag{B.6}$$

when $\beta_0 = 0$, as in (B.5). Notationally, $p_{11} = \zeta_3$, $p_{10} = \zeta_2$, and $p_{01} = \zeta_1 - \zeta_3$. The Jensen's inequality is strict if there exist $(y, d, z) \in \{0, 1\}^3$ such that

$$\frac{p_{yd,z}(\psi, \pi)}{p_{yd,z}(\psi_0, \pi)} \neq 1.$$

Under TC2, this condition can be readily shown to hold by a slight modification of the identification proof of Theorem 4.1 in HV16, which is omitted here for brevity. ■

Verification of S2(iv): For $\theta \neq \theta_0$,

$$\begin{aligned}
Q_0(\theta) - Q_0(\theta_0) &= - \sum_{y,d,z=0,1} p_{yd,z}(\theta_0) \phi_{z,0} \log \frac{p_{yd,z}(\theta)}{p_{yd,z}(\theta_0)} \\
&> - \log \sum_{y,d,z=0,1} p_{yd,z}(\theta) \phi_{z,0} \\
&= 0,
\end{aligned}$$

where the Jensen's inequality is strict because there exist $(y, d, z) \in \{0, 1\}^3$ such that

$$\frac{p_{yd,z}(\theta)}{p_{yd,z}(\theta_0)} \neq 1$$

by Theorem 4.1 in HV16 under TC2. ■

Verification of S2(v): By TC3(i), Π is compact and the parameter space is the same before and after the transformation. Also, $\Theta = \bar{h}^{-1}(\boldsymbol{\Theta})$ is compact since $\boldsymbol{\Theta}$ is compact and Assumption H(i) holds. For compactness of $\Psi(\pi)$, first note that, for a given $\pi \in \Pi$, $\bar{h}_{-\pi}(\cdot, \pi)$, which is $\bar{h}(\cdot, \pi)$ except the last element, is a homeomorphism. This is because $\boldsymbol{\Theta}_{-\pi}$ is simply connected, $\bar{h}_{-\pi}(\cdot, \pi)$ is continuous, and $\Psi(\pi)$ is bounded since Θ is bounded. Then,

$$\boldsymbol{\Theta}_{-\pi} = \boldsymbol{\Theta}_{-\pi}(\pi) \equiv \bar{h}_{-\pi}(\Psi(\pi), \pi)$$

where the first equality is because the dependence parameter π does not restrict the space of the remaining elements of $\boldsymbol{\theta}$ (or by TC3(i)), and thus $\Psi(\pi) = \bar{h}_{-\pi}^{-1}(\boldsymbol{\Theta}_{-\pi}, \pi)$. Therefore $\Psi(\pi)$ is compact since $\boldsymbol{\Theta}_{-\pi}$ is compact and $\bar{h}_{-\pi}(\cdot, \pi)$ is proper. ■

Verification of S2(vi): The space of $\psi = (\beta, \zeta)$ is continuous in π since $\Psi(\pi) = \bar{h}_{-\pi}^{-1}(\boldsymbol{\Theta}_{-\pi}, \pi)$,

where $\bar{h}_{-\pi}^{-1}(\boldsymbol{\theta}_{-\pi}, \pi)$ is continuous in π by (3.12) and TC2(iii). ∎

Let $\rho_\theta(w, \theta)$ and $\rho_{\theta\theta}(w, \theta)$ denote the first and second order partial derivatives of $\rho(w, \theta)$ w.r.t. θ, respectively. Also, let $\rho_\psi(w, \theta)$ and $\rho_{\psi\psi}(w, \theta)$ denote the first and second order partial derivatives of $\rho(w, \theta)$ w.r.t. ψ, respectively. Recall

$$B(\beta) \equiv \begin{bmatrix} I_{d_\psi} & 0_{d_\psi \times 1} \\ 0_{1 \times d_\psi} & \beta \end{bmatrix} \in \mathbb{R}^{d_\theta \times d_\theta}.$$

For $\beta \neq 0$, let

$$\begin{aligned} B^{-1}(\beta)\rho_\theta(w, \theta) &\equiv \rho_\theta^\dagger(w, \theta), \\ B^{-1}(\beta)\rho_{\theta\theta}(w, \theta)B^{-1}(\beta) &\equiv \rho_{\theta\theta}^\dagger(w, \theta) + r(w, \theta), \end{aligned} \tag{B.7}$$

where $\rho_{\theta\theta}^\dagger(w, \theta)$ is symmetric and $\rho_\theta^\dagger(w, \theta)$, $\rho_{\theta\theta}^\dagger(w, \theta)$, and $r(w, \theta)$ satisfy Assumption S3 below[30]; see below for actual expressions of these terms. Next, define

$$V^\dagger(\theta_1, \theta_2; \gamma_0) \equiv Cov_{\gamma_0}\left(\rho_\theta^\dagger(W_i, \theta_1), \rho_\theta^\dagger(W_i, \theta_2)\right).$$

Let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the maximum and minimum eigenvalues, respectively, of a square matrix A.

In this example of a threshold crossing model, define $D_\theta p_{yd,z}^\dagger(\theta) \equiv B^{-1}(\beta)D_\theta p_{yd,z}(\theta)$ so that

$$\rho_\theta(w, \theta) = -\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(w) \frac{1}{p_{yd,z}(\theta)} D_\theta p_{yd,z}(\theta),$$

$$\rho_{\theta\theta}(w, \theta) = -\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(w) \left[-\frac{1}{p_{yd,z}(\theta)^2} D_\theta p_{yd,z}(\theta) D_\theta p_{yd,z}(\theta)' + \frac{1}{p_{yd,z}(\theta)} D_{\theta\theta} p_{yd,z}(\theta) \right],$$

$$\rho_\theta^\dagger(w, \theta) = -\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(w) \frac{1}{p_{yd,z}(\theta)} D_\theta p_{yd,z}^\dagger(\theta),$$

$$\rho_{\theta\theta}^\dagger(w, \theta) = \rho_\theta^\dagger(w, \theta)\rho_\theta^\dagger(w, \theta)' = \sum_{y,d,z=0,1} \mathbf{1}_{ydz}(w) \frac{1}{p_{yd,z}(\theta)^2} D_\theta p_{yd,z}^\dagger(\theta) D_\theta p_{yd,z}^\dagger(\theta)',$$

$$r(w, \theta) = -\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(w) \frac{1}{p_{yd,z}(\theta)} B^{-1}(\beta) D_{\theta\theta} p_{yd,z}(\theta) B^{-1}(\beta).$$

Suppressing the argument (ζ_1, ζ_3, π) in h_3 and its derivatives, and suppressing the argument

---

[30]The remainder term $r(w, \theta)$ and related conditions in S3 are slightly more general than conditions on $\beta^{-1}\varepsilon(w, \theta)$ and related conditions in AC13.

(ζ_1, ζ_2, π) in h_2 and its derivatives, note that from (7.1)–(7.2),

$$
D_\theta p_{11,0}(\theta) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad
D_\theta p_{10,0}(\theta) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad
D_\theta p_{01,0}(\theta) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \quad
D_\theta p_{00,0}(\theta) = \begin{bmatrix} 0 \\ -1 \\ -1 \\ 0 \\ 0 \end{bmatrix},
$$

$$
\begin{aligned}
D_\theta p_{11,1}(\theta) &= \begin{bmatrix} C_2\left(h_3, \zeta_1 + \beta; \pi\right) \\ C_2\left(h_3, \zeta_1 + \beta; \pi\right) + C_1\left(h_3, \zeta_1 + \beta; \pi\right) h_{3,\zeta_1} \\ 0 \\ C_1\left(h_3, \zeta_1 + \beta; \pi\right) h_{3,\zeta_3} \\ C_\pi\left(h_3, \zeta_1 + \beta; \pi\right) + C_1\left(h_3, \zeta_1 + \beta; \pi\right) h_{3,\pi} \end{bmatrix} \\
&= \begin{bmatrix} C_2\left(h_3, \zeta_1 + \beta; \pi\right) \\ C_2\left(h_3, \zeta_1 + \beta; \pi\right) + C_1\left(h_3, \zeta_1 + \beta; \pi\right) h_{3,\zeta_1} \\ 0 \\ C_1\left(h_3, \zeta_1 + \beta; \pi\right) h_{3,\zeta_3} \\ \beta\left\{C_{\pi 2}\left(h_3, \zeta_1 + \beta^\dagger; \pi\right) + C_{12}\left(h_3, \zeta_1 + \beta^\dagger; \pi\right) h_{3,\pi}\right\} \end{bmatrix}, \quad \text{(B.8)}
\end{aligned}
$$

where $0 \leq |\beta^\dagger| \leq \beta$. The last equality is derived using a mean value expansion and the fact that $C_\pi\left(h_3, \zeta_1; \pi\right) + C_1\left(h_3, \zeta_1; \pi\right) h_{3,\pi} = 0$, obtained by differentiating $C(h_3, \zeta_1; \pi) = \zeta_3$ w.r.t. π. Furthermore,

$$
\begin{aligned}
D_\theta p_{10,1}(\theta) &= \begin{bmatrix} -C_2\left(h_2, \zeta_1 + \beta; \pi\right) \\ h_{2,\zeta_1} - C_2\left(h_2, \zeta_1 + \beta; \pi\right) - C_1\left(h_2, \zeta_1 + \beta; \pi\right) h_{2,\zeta_1} \\ h_{2,\zeta_2} - C_1\left(h_2, \zeta_1 + \beta; \pi\right) h_{2,\zeta_2} \\ 0 \\ h_{2,\pi} - C_\pi\left(h_2, \zeta_1 + \beta; \pi\right) - C_1\left(h_2, \zeta_1 + \beta; \pi\right) h_{2,\pi} \end{bmatrix} \\
&= \begin{bmatrix} -C_2\left(h_2, \zeta_1 + \beta; \pi\right) \\ h_{2,\zeta_1} - C_2\left(h_2, \zeta_1 + \beta; \pi\right) - C_1\left(h_2, \zeta_1 + \beta; \pi\right) h_{2,\zeta_1} \\ h_{2,\zeta_2} - C_1\left(h_2, \zeta_1 + \beta; \pi\right) h_{2,\zeta_2} \\ 0 \\ -\beta\left\{C_{\pi 2}\left(h_2, \zeta_1 + \beta^{\dagger\dagger}; \pi\right) + C_{12}\left(h_2, \zeta_1 + \beta^{\dagger\dagger}; \pi\right) h_{2,\pi}\right\} \end{bmatrix}, \quad \text{(B.9)}
\end{aligned}
$$

where $0 \leq |\beta^{\dagger\dagger}| \leq \beta$ and the last equality is derived using a mean value expansion and the fact

55

that $h_{2,\pi} - C_\pi(h_2, \zeta_1; \pi) - C_1(h_2, \zeta_1; \pi) h_{2,\pi} = 0$. Finally,

$$D_\theta p_{01,1}(\theta) = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} - D_\theta p_{11,1}(\theta), \quad D_\theta p_{00,1}(\theta) = \begin{bmatrix} -1 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} - D_\theta p_{10,1}(\theta).$$

Also, note that for all (y, d),

$$D_{\theta\theta} p_{yd,0}(\theta) = 0 \tag{B.10}$$

and

$$D_{\theta\theta} p_{01,1}(\theta) = -D_{\theta\theta} p_{11,1}(\theta), \qquad D_{\theta\theta} p_{00,1}(\theta) = -D_{\theta\theta} p_{10,1}(\theta). \tag{B.11}$$

Now, for $z = 0$,

$$D_\theta p_{yd,z}^\dagger(\theta) = D_\theta p_{yd,z}(\theta) \tag{B.12}$$

and, for $z = 1$,

$$D_\theta p_{11,1}^\dagger(\theta) = \begin{bmatrix} C_2(h_3, \zeta_1 + \beta; \pi) \\ C_2(h_3, \zeta_1 + \beta; \pi) + C_1(h_3, \zeta_1 + \beta; \pi) h_{3,\zeta_1} \\ 0 \\ C_1(h_3, \zeta_1 + \beta; \pi) h_{3,\zeta_3} \\ C_{\pi 2}(h_3, \zeta_1 + \beta^\dagger; \pi) + C_{12}(h_3, \zeta_1 + \beta^\dagger; \pi) h_{3,\pi} \end{bmatrix}, \tag{B.13}$$

$$D_\theta p_{10,1}^\dagger(\theta) = \begin{bmatrix} -C_2(h_2, \zeta_1 + \beta; \pi) \\ h_{2,\zeta_1} - C_2(h_2, \zeta_1 + \beta; \pi) - C_1(h_2, \zeta_1 + \beta; \pi) h_{2,\zeta_1} \\ h_{2,\zeta_2} - C_1(h_2, \zeta_1 + \beta; \pi) h_{2,\zeta_2} \\ 0 \\ -C_{\pi 2}(h_2, \zeta_1 + \beta^{\dagger\dagger}; \pi) - C_{12}(h_2, \zeta_1 + \beta^{\dagger\dagger}; \pi) h_{2,\pi} \end{bmatrix}, \tag{B.14}$$

and expressions for the remaining two derivatives can be derived analogously.

Note that

$$\rho_\psi(w, \theta) = -\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(w) \frac{1}{p_{yd,z}(\theta)} D_\psi p_{yd,z}(\theta),$$

56

$$\rho_{\psi\psi}(w,\theta) = -\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(w) \left[-\frac{1}{p_{yd,z}(\theta)^2} D_\psi p_{yd,z}(\theta) D_\psi p_{yd,z}(\theta)' + \frac{1}{p_{yd,z}(\theta)} D_{\psi\psi} p_{yd,z}(\theta) \right],$$

where, with $\psi = (\beta, \zeta) = (\beta, \zeta_1, \zeta_2, \zeta_3)$,

$$D_\psi p_{11,0}(\theta) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad D_\psi p_{10,0}(\theta) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad D_\psi p_{01,0}(\theta) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \end{bmatrix}, \quad D_\psi p_{00,0}(\theta) = \begin{bmatrix} 0 \\ -1 \\ -1 \\ 0 \end{bmatrix},$$

$$D_\psi p_{11,1}(\theta) = \begin{bmatrix} C_2\left(h_3, \zeta_1 + \beta; \pi\right) \\ C_2\left(h_3, \zeta_1 + \beta; \pi\right) + C_1\left(h_3, \zeta_1 + \beta; \pi\right) h_{3,\zeta_1} \\ 0 \\ C_1\left(h_3, \zeta_1 + \beta; \pi\right) h_{3,\zeta_3} \end{bmatrix},$$

$$D_\psi p_{10,1}(\theta) = \begin{bmatrix} -C_2\left(h_2, \zeta_1 + \beta; \pi\right) \\ h_{2,\zeta_1} - C_2\left(h_2, \zeta_1 + \beta; \pi\right) - C_1\left(h_2, \zeta_1 + \beta; \pi\right) h_{2,\zeta_1} \\ h_{2,\zeta_2} - C_1\left(h_2, \zeta_1 + \beta; \pi\right) h_{2,\zeta_2} \\ 0 \end{bmatrix},$$

and

$$D_\psi p_{01,1}(\theta) = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} - D_\psi p_{11,1}(\theta), \quad D_\psi p_{00,1}(\theta) = \begin{bmatrix} -1 \\ -1 \\ 0 \\ 0 \end{bmatrix} - D_\psi p_{10,1}(\theta).$$

Also, for all (y, d) and θ,

$$D_{\psi\psi} p_{yd,0}(\theta) = 0 \tag{B.15}$$

and

$$D_{\psi\psi} p_{01,1}(\theta) = -D_{\psi\psi} p_{11,1}(\theta), \qquad D_{\psi\psi} p_{00,1}(\theta) = -D_{\psi\psi} p_{10,1}(\theta). \tag{B.16}$$

Assumption S3: (i) (a) $E_{\gamma_0} r(W_i, \theta_0) = 0$; and (b) $\|E_{\gamma_0} r(W_i, \psi_0, \pi)\| \leq C |\pi - \pi_0| \; \forall \gamma_0 \in \Gamma$ with $0 < |\beta_0| < \delta$ for some $\delta > 0$.
(ii) (a) For all $\delta > 0$ and some function $M_1(w) : \mathcal{W} \to \mathbb{R}_+$, $\|\rho_{\psi\psi}(w, \theta_1) - \rho_{\psi\psi}(w, \theta_2)\| + \left\|\rho_{\theta\theta}^\dagger(w, \theta_1) - \rho_{\theta\theta}^\dagger(w, \theta_2)\right\| \leq M_1(w)\delta$, $\forall \theta_1, \theta_2 \in \Theta$ with $\|\theta_1 - \theta_2\| \leq \delta$, $\forall w \in \mathcal{W}$; and (b) for all $\delta > 0$ and some function $M_2(w) : \mathcal{W} \to \mathbb{R}_+$, $\left\|\rho_\theta^\dagger(w, \theta_1) - \rho_\theta^\dagger(w, \theta_2)\right\| + \|r(w, \theta_1) - r(w, \theta_2)\| \leq$

$M_2(w)\delta$, $\forall \theta_1, \theta_2 \in \Theta$ with $\|\theta_1 - \theta_2\| \leq \delta$, $\forall w \in \mathcal{W}$.

(iii) $E_{\gamma_0} \sup_{\theta \in \Theta} \left\{ |\rho(W_i, \theta)|^{1+\delta} + \|\rho_{\psi\psi}(W_i, \theta)\|^{1+\delta} + \left\|\rho_{\theta\theta}^{\dagger}(W_i, \theta)\right\|^{1+\delta} + M_1(W_i) + \left\|\rho_{\theta}^{\dagger}(W_i, \theta)\right\|^q + \|r(W_i, \theta)\|^q + M_2(W_i)^q \leq C \right.$ for some $\delta > 0$ $\forall \gamma_0 \in \Gamma$, where q is as in Assumption S1.

(iv) (a) $\lambda_{\min}(E_{\gamma_0}\rho_{\psi\psi}(W_i, \psi_0, \pi)) > 0$ $\forall \pi \in \Pi$ when $\beta_0 = 0$; and (b) $E_{\gamma_0}\rho_{\theta\theta}^{\dagger}(W_i, \theta_0)$ is positive definite $\forall \gamma_0 \in \Gamma$.

(v) $V^{\dagger}(\theta_0, \theta_0; \gamma_0)$ is positive definite $\forall \gamma_0 \in \Gamma$.

Verification of S3(i)(a): Note that

$$E_{\gamma_0} r(W_i, \theta_0) = -\sum_{y,d,z=0,1} \phi_{z,0} B^{-1}(\beta_0) D_{\theta\theta} p_{yd,z}(\theta_0) B^{-1}(\beta_0) = 0$$

by (B.10) and (B.11) since $\beta_0 \neq 0$. ∎

Verification of S3(i)(b): Using (B.10) and (B.11),

$$E_{\gamma_0} r(W_i, \psi_0, \pi) = \sum_{y,d=0,1} p_{yd,1}(\theta_0) \phi_{1,0} B^{-1}(\beta_0) \frac{D_{\theta\theta} p_{yd,1}(\psi_0, \pi)}{p_{yd,1}(\psi_0, \pi)} B^{-1}(\beta_0)$$

$$= \phi_{1,0} B^{-1}(\beta_0) \left[\frac{p_{11,1}(\theta_0)}{p_{11,1}(\psi_0, \pi)} D_{\theta\theta} p_{11,1}(\psi_0, \pi) + \frac{p_{01,1}(\theta_0)}{p_{01,1}(\psi_0, \pi)} D_{\theta\theta} p_{01,1}(\psi_0, \pi) \right.$$

$$\left. + \frac{p_{10,1}(\theta_0)}{p_{10,1}(\psi_0, \pi)} D_{\theta\theta} p_{10,1}(\psi_0, \pi) + \frac{p_{00,1}(\theta_0)}{p_{00,1}(\psi_0, \pi)} D_{\theta\theta} p_{00,1}(\psi_0, \pi) \right] B^{-1}(\beta_0)$$

$$= \phi_{1,0} B^{-1}(\beta_0) \left[\left(\frac{p_{11,1}(\theta_0)}{p_{11,1}(\psi_0, \pi)} - \frac{p_{01,1}(\theta_0)}{p_{01,1}(\psi_0, \pi)} \right) D_{\theta\theta} p_{11,1}(\psi_0, \pi) \right.$$

$$\left. + \left(\frac{p_{10,1}(\theta_0)}{p_{10,1}(\psi_0, \pi)} - \frac{p_{00,1}(\theta_0)}{p_{00,1}(\psi_0, \pi)} \right) D_{\theta\theta} p_{10,1}(\psi_0, \pi) \right] B^{-1}(\beta_0)$$

$$= \phi_{1,0} B^{-1}(\beta_0) \left[\left(\frac{(\zeta_{10} + \beta_0)(p_{11,1}(\theta_0) - p_{11,1}(\psi_0, \pi))}{p_{11,1}(\psi_0, \pi)(\zeta_{10} + \beta_0 - p_{11,1}(\psi_0, \pi))} \right) D_{\theta\theta} p_{11,1}(\psi_0, \pi) \right.$$

$$\left. + \left(\frac{(1 - \zeta_{10} - \beta_0)(p_{10,1}(\theta_0) - p_{10,1}(\psi_0, \pi))}{p_{10,1}(\psi_0, \pi)(1 - \zeta_{10} - \beta_0 - p_{10,1}(\psi_0, \pi))} \right) D_{\theta\theta} p_{10,1}(\psi_0, \pi) \right] B^{-1}(\beta_0) \quad \text{(B.17)}$$

where the last equality uses $p_{01,1}(\theta) = \zeta_1 + \beta - p_{11,1}(\theta)$ and $p_{00,1}(\theta) = 1 - \zeta_1 - \beta - p_{10,1}(\theta)$. Apply the mean value theorem to $p_{11,1}(\theta_0) - p_{11,1}(\psi_0, \pi)$ w.r.t. π:

$$p_{11,1}(\psi_0, \pi_0) - p_{11,1}(\psi_0, \pi) = \frac{\partial p_{11,1}(\psi_0, \pi^{\dagger})}{\partial \pi}(\pi_0 - \pi)$$

$$= \frac{\partial^2 p_{11,1}(\beta^{\dagger}, \zeta_0, \pi^{\dagger})}{\partial \pi \partial \beta}(\pi_0 - \pi)\beta_0, \quad \text{(B.18)}$$

where $\pi^{\dagger}$ is between π_0 and π and $0 \leq |\beta^{\dagger}| \leq |\beta_0|$. The second equality holds by another mean

value expansion of $\frac{\partial p_{11,1}(\psi_0, \pi^\dagger)}{\partial \pi}$ w.r.t. β_0 around $\beta_0 = 0$ and the fact that $\left.\frac{\partial p_{11,1}(\beta, \zeta_0, \pi^\dagger)}{\partial \pi}\right|_{\beta=0} = 0$ since

$$C_\pi(h_3(\pi), \zeta_1; \pi) + C_1(h_3(\pi), \zeta_1; \pi) h_{3,\pi}(\pi) = 0$$

for all (ζ_1, ζ_3, π). Similarly, using mean value expansions,

$$p_{10,1}(\psi_0, \pi_0) - p_{10,1}(\psi_0, \pi) = \frac{\partial^2 p_{10,1}(\beta^{\dagger\dagger}, \zeta_0, \pi^{\dagger\dagger})}{\partial \pi \partial \beta}(\pi_0 - \pi)\beta_0 \tag{B.19}$$

for some $\pi^{\dagger\dagger}$ between π_0 and π and $0 \leq |\beta^{\dagger\dagger}| \leq |\beta_0|$. Therefore, combining (B.17)–(B.19),

$$\begin{aligned}
\|E_{\gamma_0} r(W_i, \psi_0, \pi)\| &\leq |c_1| \left\|B^{-1}(\beta_0)\beta_0 D_{\theta\theta} p_{11,1}(\psi_0, \pi) B^{-1}(\beta_0)\right\| |\pi_0 - \pi| \\
&\quad + |c_2| \left\|B^{-1}(\beta_0)\beta_0 D_{\theta\theta} p_{10,1}(\psi_0, \pi) B^{-1}(\beta_0)\right\| |\pi_0 - \pi|
\end{aligned}$$

where c_1 and c_2 are collections of all other terms, whose norms are bounded by (7.1)–(7.2) and Lemma B.1. Also $\left\|B^{-1}(\beta_0)\beta_0\right\|$ is bounded for $0 < |\beta_0| < \delta$. Note that $\left\|D_{\theta\theta} p_{11,1}(\psi_0, \pi) B^{-1}(\beta_0)\right\|$ and $\left\|D_{\theta\theta} p_{10,1}(\psi_0, \pi) B^{-1}(\beta_0)\right\|$ can be shown to be bounded for $0 < |\beta_0| < \delta$ by differentiating (B.13) and (B.14) w.r.t. θ, respectively, and applying Lemma B.1. ∎

Verification of S3(ii)(a): Generically, for $A = aa'$ where $a = (a_1, ..., a_p) \in \mathbb{R}^{d_a}$ and $a_1, ..., a_p$ are vectors,

$$\|A\| \leq \sum_{j=1}^{p} \|a_j\|^2,$$

and for $A^* = a^* a^{*\prime}$

$$\begin{aligned}
\|A - A^*\| &\leq \left\|a(a - a^*)'\right\| + \left\|(a - a^*)a^{*\prime}\right\| \leq (\|a\| + \|a^*\|) \|a - a^*\| \\
&\leq \sum_{j=1}^{p} \left(\|a_j\| + \|a_j^*\|\right) \sum_{j=1}^{p} \left\|a_j - a_j^*\right\|.
\end{aligned}$$

Applying this result to the last inequality below,

$$\begin{aligned}
&\|\rho_{\psi\psi}(w, \theta_1) - \rho_{\psi\psi}(w, \theta_2)\| \\
&\leq \sum_{y,d,z=0,1} \left\| \frac{D_\psi p_{yd,z}(\theta_1) D_\psi p_{yd,z}(\theta_1)'}{p_{yd,z}(\theta_1)^2} - \frac{D_\psi p_{yd,z}(\theta_2) D_\psi p_{yd,z}(\theta_2)'}{p_{yd,z}(\theta_2)^2} \right\|
\end{aligned}$$

$$+ \sum_{y,d,z=0,1} \left\| \frac{D_{\psi\psi}p_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)} - \frac{D_{\psi\psi}p_{yd,z}(\theta_2)}{p_{yd,z}(\theta_2)} \right\|$$

$$\leq \sum_{y,d,z=0,1} \left(\left\| \frac{D_{\psi}p_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)} \right\| + \left\| \frac{D_{\psi}p_{yd,z}(\theta_2)}{p_{yd,z}(\theta_2)} \right\| \right) \left\| \frac{D_{\psi}p_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)} - \frac{D_{\psi}p_{yd,z}(\theta_2)}{p_{yd,z}(\theta_2)} \right\|$$

$$+ \sum_{y,d,z=0,1} \left\| \frac{D_{\psi\psi}p_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)} - \frac{D_{\psi\psi}p_{yd,z}(\theta_2)}{p_{yd,z}(\theta_2)} \right\|$$

$$\leq \sum_{y,d,z=0,1} \sum_{j=1}^{d_\psi} \left(\left| \frac{D_{\psi_j}p_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)} \right| + \left| \frac{D_{\psi_j}p_{yd,z}(\theta_2)}{p_{yd,z}(\theta_2)} \right| \right) \sum_{j=1}^{d_\psi} \left| \frac{D_{\psi_j}p_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)} - \frac{D_{\psi_j}p_{yd,z}(\theta_2)}{p_{yd,z}(\theta_2)} \right|$$

$$+ \sum_{y,d,z=0,1} \sum_{j,k=1}^{d_\psi} \left| \frac{D_{\psi_j\psi_k}p_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)} - \frac{D_{\psi_j\psi_k}p_{yd,z}(\theta_2)}{p_{yd,z}(\theta_2)} \right|,$$

where $|\mathbf{1}_{ydz}(w)| \leq 1$ is used in the first inequality. Applying the mean value theorem to the differential terms,

$$\left| \frac{D_{\psi_j}p_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)} - \frac{D_{\psi_j}p_{yd,z}(\theta_2)}{p_{yd,z}(\theta_2)} \right| \leq \left\| D_\theta \left\{ \frac{D_{\psi_j}p_{yd,z}(\theta^\dagger)}{p_{yd,z}(\theta^\dagger)} \right\} \right\| \|\theta_1 - \theta_2\|,$$

$$\left| \frac{D_{\psi_j\psi_k}p_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)} - \frac{D_{\psi_j\psi_k}p_{yd,z}(\theta_2)}{p_{yd,z}(\theta_2)} \right| \leq \left\| D_\theta \left\{ \frac{D_{\psi_j\psi_k}p_{yd,z}(\theta^{\dagger\dagger})}{p_{yd,z}(\theta^{\dagger\dagger})} \right\} \right\| \|\theta_1 - \theta_2\|,$$

where $\theta^\dagger$ and $\theta^{\dagger\dagger}$ lie between θ_1 and θ_2 (element-wise). By Lemma B.1, $\sup_\theta \left| \frac{D_{\psi_j}p_{yd,z}(\theta)}{p_{yd,z}(\theta)} \right| < c_1$, $\sup_\theta \left| D_{\theta_k} \left\{ \frac{D_{\psi_j}p_{yd,z}(\theta)}{p_{yd,z}(\theta)} \right\} \right| < c_2$ and $\sup_\theta \left| D_{\theta_l} \left\{ \frac{D_{\psi_j\psi_k}p_{yd,z}(\theta)}{p_{yd,z}(\theta)} \right\} \right| < c_3$ for some positive constants c_1, c_2 and c_3, and therefore combining the inequalities,

$$\|\rho_{\psi\psi}(w,\theta_1) - \rho_{\psi\psi}(w,\theta_2)\| \leq \sum_{y,d,z=0,1} \sum_{j=1}^{d_\psi} 2c_1 \sum_{j=1}^{d_\psi} \sum_{k=1}^{d_\theta} c_2 \|\theta_1 - \theta_2\|$$

$$+ \sum_{y,d,z=0,1} \sum_{j,k=1}^{d_\psi} \sum_{l=1}^{d_\theta} c_3 \|\theta_1 - \theta_2\|. \tag{B.20}$$

Similarly,

$$\left\| \rho_{\theta\theta}^\dagger(w,\theta_1) - \rho_{\theta\theta}^\dagger(w,\theta_2) \right\|$$

$$\leq \sum_{y,d,z=0,1} \left\| \frac{D_\theta p_{yd,z}^\dagger(\theta_1) D_\theta p_{yd,z}^\dagger(\theta_1)'}{p_{yd,z}(\theta_1)^2} - \frac{D_\theta p_{yd,z}^\dagger(\theta_2) D_\theta p_{yd,z}^\dagger(\theta_2)'}{p_{yd,z}(\theta_2)^2} \right\|$$

$$\leq \sum_{y,d,z=0,1} \left(\left\| \frac{D_\theta p^\dagger_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)} \right\| + \left\| \frac{D_\theta p^\dagger_{yd,z}(\theta_2)}{p_{yd,z}(\theta_2)} \right\| \right) \left\| \frac{D_\theta p^\dagger_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)} - \frac{D_\theta p^\dagger_{yd,z}(\theta_2)}{p_{yd,z}(\theta_2)} \right\|$$

$$\leq \sum_{y,d,z=0,1} \sum_{j=1}^{d_\theta} \left(\left| \frac{D_{\theta_j} p^\dagger_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)} \right| + \left| \frac{D_{\theta_j} p^\dagger_{yd,z}(\theta_2)}{p_{yd,z}(\theta_2)} \right| \right) \sum_{j=1}^{d_\theta} \left| \frac{D_{\theta_j} p^\dagger_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)} - \frac{D_{\theta_j} p^\dagger_{yd,z}(\theta_2)}{p_{yd,z}(\theta_2)} \right|$$

and by Lemma B.1, $\sup_\theta \left| \frac{D_{\theta_j} p^\dagger_{yd,z}(\theta)}{p_{yd,z}(\theta)} \right| < c_4$ and $\sup_\theta \left| D_{\theta_k} \left\{ \frac{D_{\theta_j} p^\dagger_{yd,z}(\theta)}{p_{yd,z}(\theta)} \right\} \right| < c_5$ for some positive constants c_4 and c_5, and therefore by applying the mean value theorem as above,

$$\left\| \rho^\dagger_{\theta\theta}(w,\theta_1) - \rho^\dagger_{\theta\theta}(w,\theta_2) \right\| \leq \sum_{y,d,z=0,1} \sum_{j=1}^{d_\theta} 2c_4 \sum_{j,k=1}^{d_\theta} c_5 \left\| \theta_1 - \theta_2 \right\|. \tag{B.21}$$

By combining (B.20) and (B.21), we have the desired result. ∎

Verification of S3(ii)(b): For bounding $\|r(w,\theta_1) - r(w,\theta_2)\|$, the proof is very similar to the one above with $\left\| \rho^\dagger_{\theta\theta}(w,\theta_1) - \rho^\dagger_{\theta\theta}(w,\theta_2) \right\|$. Bounding $\left\| \rho^\dagger_\theta(w,\theta_1) - \rho^\dagger_\theta(w,\theta_2) \right\|$ can also be done analogously. ∎

Verification of S3(iii): First, $M_1(w)$ is finite and does not depend on w, as can be seen from the verification of S3(ii)(a). Now, since $|\mathbf{1}_{ydz}(w)| \leq 1$

$$E_{\gamma_0} \sup_{\theta\in\Theta} |\rho(W_i,\theta)|^{1+\delta} \leq E_{\gamma_0} \left(\sum_{y,d,z=0,1} \sup_{\theta\in\Theta} |\mathbf{1}_{ydz}(w) \cdot \log p_{yd,z}(\theta)| \right)^{1+\delta}$$

$$\leq \left(\sum_{y,d,z=0,1} \sup_{\theta\in\Theta} |\log p_{yd,z}(\theta)| \right)^{1+\delta},$$

which is bounded since $p_{yd,z}(\theta)$ is bounded away from zero for any $\theta \in \Theta$ and $(y,d,z) \in \{0,1\}$ by Lemma B.1. Next,

$$E_{\gamma_0} \sup_{\theta\in\Theta} \|\rho_{\psi\psi}(W_i,\theta)\|^{1+\delta}$$

$$\leq E_{\gamma_0} \left(\sum_{y,d,z=0,1} \sup_{\theta\in\Theta} \left\| \mathbf{1}_{ydz}(w) \left[-\frac{1}{p_{yd,z}(\theta)^2} D_\psi p_{yd,z}(\theta) D_\psi p_{yd,z}(\theta)' + \frac{1}{p_{yd,z}(\theta)} D_{\psi\psi} p_{yd,z}(\theta) \right] \right\| \right)^{1+\delta}$$

$$\leq \left(\sum_{y,d,z=0,1} C \sup_{\theta\in\Theta} \left\{ \left\| D_\psi p_{yd,z}(\theta) D_\psi p_{yd,z}(\theta)' \right\| + \left\| D_{\psi\psi} p_{yd,z}(\theta) \right\| \right\} \right)^{1+\delta}$$

61

by Lemma B.1, where $\|D_\psi p_{yd,z}(\theta) D_\psi p_{yd,z}(\theta)'\| \leq \sum_{j=1}^{d_\psi} \|D_{\psi_j} p_{yd,z}(\theta)\|^2$, which is bounded by Lemma B.1, and similarly for $\|D_{\psi\psi} p_{yd,z}(\theta)\|$. Similar arguments to those used in the verification of S3(i)(b) and S3(ii)(a) provide the desired result for the remaining four terms in the assumption. ∎

Verification of S3(iv)(a): Note that, when $\beta_0 = 0$,

$$E_{\gamma_0} \rho_{\psi\psi}(W_i, \psi_0, \pi) = \sum_{y,d,z=0,1} p_{yd,z}(\theta_0) \phi_{z,0} \left[\frac{D_\psi p_{yd,z}(\psi_0, \pi) D_\psi p_{yd,z}(\psi_0, \pi)'}{p_{yd,z}(\psi_0, \pi)^2} - \frac{D_{\psi\psi} p_{yd,z}(\psi_0, \pi)}{p_{yd,z}(\psi_0, \pi)} \right]$$

$$= \sum_{y,d,z=0,1} \phi_{z,0} \left[\frac{D_\psi p_{yd,z}(\psi_0, \pi) D_\psi p_{yd,z}(\psi_0, \pi)'}{p_{yd}^0} - D_{\psi\psi} p_{yd,z}(\psi_0, \pi) \right]$$

$$= \sum_{y,d,z=0,1} \phi_{z,0} \frac{D_\psi p_{yd,z}(\psi_0, \pi) D_\psi p_{yd,z}(\psi_0, \pi)'}{p_{yd}^0}$$

where the second equality is by (B.6), and the third equality is by (B.15) and (B.16). Let $M_{yd,z} \equiv D_\psi p_{yd,z}(\psi_0, \pi) D_\psi p_{yd,z}(\psi_0, \pi)'$ and $\tilde{M}_{yd,z} \equiv M_{yd,z}/p_{yd}^0$ so that

$$E_{\gamma_0} \rho_{\psi\psi}(W_i, \psi_0, \pi) = \phi_{1,0} \sum_{y,d=0,1} \tilde{M}_{yd,1} + \phi_{0,0} \sum_{y,d=0,1} \tilde{M}_{yd,0}. \tag{B.22}$$

Let $h_3(\pi) \equiv h_3(\zeta_{10}, \zeta_{30}; \pi)$ and $h_2(\pi) \equiv h_2(\zeta_{10}, \zeta_{20}; \pi)$. Note that when $\beta_0 = 0$, the $D_\psi p_{yd,z}(\psi_0, \pi)$ terms can be expressed as

$$D_\psi p_{11,0} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad D_\psi p_{10,0} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad D_\psi p_{01,0} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \end{bmatrix}, \quad D_\psi p_{00,0} = \begin{bmatrix} 0 \\ -1 \\ -1 \\ 0 \end{bmatrix},$$

$$D_\psi p_{11,1} = \begin{bmatrix} C_2\left(h_3(\pi), \zeta_1; \pi\right) \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

$$D_\psi p_{10,1} = \begin{bmatrix} -C_2\left(h_2(\pi), \zeta_1; \pi\right) \\ 0 \\ 1 \\ 0 \end{bmatrix},$$

62

and

$$D_\psi p_{01,1} = \begin{bmatrix} 1 - C_2\left(h_3(\pi), \zeta_1; \pi\right) \\ 1 \\ 0 \\ -1 \end{bmatrix}, \quad D_\psi p_{00,1} = \begin{bmatrix} -1 + C_2\left(h_2(\pi), \zeta_1; \pi\right) \\ -1 \\ -1 \\ 0 \end{bmatrix},$$

where, in $D_\psi p_{11,1}$ and $D_\psi p_{10,1}$,

$$C_2\left(h_3, \zeta_1; \pi\right) + C_1\left(h_3, \zeta_1; \pi\right) h_{3,\zeta_1} = 0, \tag{B.23}$$

$$C_1\left(h_3, \zeta_1; \pi\right) h_{3,\zeta_3} = 1, \tag{B.24}$$

$$h_{2,\zeta_1} - C_2\left(h_2, \zeta_1; \pi\right) - C_1\left(h_2, \zeta_1; \pi\right) h_{2,\zeta_1} = 0, \tag{B.25}$$

$$h_{2,\zeta_2} - C_1\left(h_2, \zeta_1; \pi\right) h_{2,\zeta_2} = 1, \tag{B.26}$$

by differentiating the objects in (7.1)–(7.2) w.r.t. ζ_1, ζ_2 and ζ_3 and (B.5). Let $c \equiv C_2\left(h_3(\pi), \zeta_{10}; \pi\right)$ and $\tilde{c} \equiv C_2\left(h_2(\pi), \zeta_{10}; \pi\right)$ for notational simplicity. Then,

$$M_{11,1} = \begin{bmatrix} c^2 & 0 & 0 & c \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ c & 0 & 0 & 1 \end{bmatrix}, \quad M_{10,1} = \begin{bmatrix} \tilde{c}^2 & 0 & -\tilde{c} & 0 \\ 0 & 0 & 0 & 0 \\ -\tilde{c} & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$M_{01,1} = \begin{bmatrix} (1-c)^2 & 1-c & 0 & c-1 \\ 1-c & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ c-1 & -1 & 0 & 1 \end{bmatrix}, \quad M_{00,1} = \begin{bmatrix} (1-\tilde{c})^2 & 1-\tilde{c} & 1-\tilde{c} & 0 \\ 1-\tilde{c} & 1 & 1 & 0 \\ 1-\tilde{c} & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$M_{11,0} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad M_{10,0} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$M_{01,0} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}, \quad M_{00,0} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

By Weyl (1912),

$$\lambda_{\min}(A + B) \geq \lambda_{\min}(A) + \lambda_{\min}(B) \tag{B.27}$$

for symmetric matrices A and B. Thus, for (B.22),

$$\lambda_{\min}\left(E_{\gamma_0}\rho_{\psi\psi}(W_i, \psi_0, \pi)\right) \geq \lambda_{\min}\left(\phi_{1,0}\sum_{y,d=0,1}\tilde{M}_{yd,1}\right) + \lambda_{\min}\left(\phi_{0,0}\sum_{y,d=0,1}\tilde{M}_{yd,0}\right).$$

The second term on the right hand side satisfies $\lambda_{\min}\left(\phi_{0,0}\sum_{y,d=0,1}\tilde{M}_{yd,0}\right) \geq \phi_{0,0}\sum_{y,d=0,1}\lambda_{\min}\left(\tilde{M}_{yd,0}\right) = 0$ by (B.27), the above expressions for the $M_{yd,0}$'s and since $\lambda_{\min}\left(\tilde{M}_{yd,0}\right) = \lambda_{\min}\left(M_{yd,0}\right) = 0$ because $p_{yd}^0 > 0$ for all (y,d) by Lemma B.1(v). The first term on the right hand side satisfies $\lambda_{\min}\left(\phi_{1,0}\sum_{y,d=0,1}\tilde{M}_{yd,1}\right) \geq \phi_{1,0}\lambda_{\min}\left(\left\{\tilde{M}_{11,1} + \tilde{M}_{01,1} + \tilde{M}_{00,1}\right\}\right)$ by (B.27) and since $\lambda_{\min}\left(\tilde{M}_{10,1}\right) = \lambda_{\min}\left(M_{10,1}\right) = 0$. Now we prove $\lambda_{\min}(\tilde{M}_{11,1} + \tilde{M}_{01,1} + \tilde{M}_{00,1}) > 0$, which then implies that $\lambda_{\min}(E_{\gamma_0}\rho_{\psi\psi}(W_i, \psi_0, \pi)) > 0$ as desired since $\phi_{1,0} > 0$ by TC5(ii). Under TC5(i) and by Lemma B.1(v), let $a \equiv p_{11}^0/p_{01}^0$ and $b \equiv p_{11}^0/p_{00}^0$ for simplicity. Then, $\tilde{M}_{11,1} + \tilde{M}_{01,1} + \tilde{M}_{00,1} = (M_{11,1} + aM_{01,1} + bM_{00,1})/p_{11}^0$ and

$$
\begin{aligned}
M &\equiv M_{11,1} + aM_{01,1} + bM_{00,1} \\
&= \begin{bmatrix}
a(1-c)^2 + b(1-\tilde{c})^2 + c^2 & a(1-c) + b(1-\tilde{c}) & b(1-\tilde{c}) & -a(1-c) + c \\
a(1-c) + b(1-\tilde{c}) & a+b & b & -a \\
b(1-\tilde{c}) & b & b & 0 \\
-a(1-c) + c & -a & 0 & a+1
\end{bmatrix}.
\end{aligned}
$$

Then one can easily show the following: For the k-th leading principal minor $|M_k|$ and determinant $|M|$ of M,

$$|M_1| = a(1-c)^2 + b(1-\tilde{c})^2 + c^2 > 0,$$

$$|M_2| = ab\left[(1-c) + (1-\tilde{c})\right]^2 + (a+b)c^2 > 0,$$

$$|M_3| = ab\tilde{c}^2 > 0,$$

$$|M| = ab\left[a(2c-1)^2 + b(\tilde{c}-1)^2\right] > 0,$$

and therefore M is positive definite and so is M/p_{11}^0, i.e., $\lambda_{\min}(\tilde{M}_{11,1} + \tilde{M}_{01,1} + \tilde{M}_{00,1}) > 0$. ∎

Verification of S3(iv)(b): We divide this proof into two cases: (i) $\beta_0 \neq 0$ and (ii) $\beta_0 = 0$.

Case (i): Note that by S3(i)(a),

$$E_{\gamma_0} B^{-1}(\beta_0) \rho_{\theta\theta}(W_i, \theta_0) B^{-1}(\beta_0) = E_{\gamma_0} \rho_{\theta\theta}^\dagger(W_i, \theta_0).$$

We first show that $E_{\gamma_0} \rho_{\theta\theta}(w, \theta_0)$ is positive definite For a positive definite matrix A, $P'AP$ is also positive definite, provided that P has full rank. Therefore, given Remark 3.1, since the full vector Jacobian $\frac{\partial \mathbf{g}(\theta_0)}{\partial \theta'}$ has full rank by HV16, it suffices to show that $\mathcal{I}^\dagger(\mathbf{g}(\theta_0))$ is positive definite where $\mathbf{g}(\theta_0) \equiv \mathbf{g}(\bar{h}(\theta_0))$. Since

$$\frac{\partial \log f^\dagger(w; \mathbf{g}(\theta_0))}{\partial \mathbf{g}'} = \left[\frac{\mathbf{1}_{110}(w)}{p_{11,0}(\theta_0)}, \frac{\mathbf{1}_{111}(w)}{p_{11,1}(\theta_0)}, ..., \frac{\mathbf{1}_{011}(w)}{p_{01,1}(\theta_0)} \right],$$

we have a diagonal matrix

$$\mathcal{I}^\dagger(\mathbf{g}(\theta_0)) = \begin{bmatrix} \frac{\phi_{0,0}}{p_{11,0}(\theta_0)} & 0 & 0 & 0 \\ 0 & \frac{\phi_{1,0}}{p_{11,1}(\theta_0)} & 0 & 0 \\ & & \ddots & \\ 0 & 0 & 0 & \frac{\phi_{1,0}}{p_{01,1}(\theta_0)} \end{bmatrix},$$

which is positive definite, since all diagonal elements are positive by Lemma B.1. Therefore $E_{\gamma_0} \rho_{\theta\theta}(w, \theta_0)$ is positive definite Thus, for a nonzero vector $a \in \mathbb{R}^{d_\theta}$, $a' E_{\gamma_0} \rho_{\theta\theta}(w, \theta_0) a > 0$, which implies that, for a nonzero vector $\tilde{a} \in \mathbb{R}^{d_\theta}$, $\tilde{a}' E_{\gamma_0} \rho_{\theta\theta}^\dagger(w, \theta_0) \tilde{a} = \tilde{a}' B^{-1}(\beta_0) E_{\gamma_0} \rho_{\theta\theta}(w, \theta_0) B^{-1}(\beta_0) \tilde{a} > 0$. Therefore $E_{\gamma_0} \rho_{\theta\theta}^\dagger(w, \theta_0)$ is positive definite.

Case (ii): First note that by (B.12)–(B.14) and (B.23)–(B.26), we can express $D_\theta p_{yd,z}^\dagger(\psi_0, \pi)$'s as follows when $\beta_0 = 0$,

$$D_\theta p_{11,0}^\dagger = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad D_\theta p_{10,0}^\dagger = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad D_\theta p_{01,0}^\dagger = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \quad D_\theta p_{00,0}^\dagger = \begin{bmatrix} 0 \\ -1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \qquad \text{(B.28)}$$

$$D_\theta p_{11,1}^\dagger = \begin{bmatrix} C_2\left(h_3(\pi), \zeta_{10}; \pi\right) \\ 0 \\ 0 \\ 1 \\ C_{\pi 2}\left(h_3(\pi), \zeta_{10}; \pi\right) + C_{12}\left(h_3(\pi), \zeta_{10}; \pi\right) h_{3,\pi}(\zeta_{10}, \zeta_{30}, \pi) \end{bmatrix}, \qquad \text{(B.29)}$$

$$D_\theta p_{10,1}^\dagger = \begin{bmatrix} -C_2\left(h_2(\pi),\zeta_{10};\pi\right) \\ 0 \\ 1 \\ 0 \\ -C_{\pi 2}\left(h_2(\pi),\zeta_{10};\pi\right) - C_{12}\left(h_2(\pi),\zeta_{10};\pi\right)h_{2,\pi}(\zeta_{10},\zeta_{20},\pi) \end{bmatrix}, \tag{B.30}$$

and

$$D_\theta p_{01,1}^\dagger = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} - D_\theta p_{11,1}^\dagger, \quad D_\theta p_{00,1}^\dagger = \begin{bmatrix} -1 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} - D_\theta p_{10,1}^\dagger. \tag{B.31}$$

The remaining arguments are similar to those used to verify S3(iv)(a): Let $M_{yd,z}^\dagger \equiv D_\theta p_{yd,z}^\dagger(\theta_0)$ $\times D_\theta p_{yd,z}^\dagger(\theta_0)'$ and $\tilde{M}_{yd,z}^\dagger \equiv M_{yd,z}^\dagger/p_{yd}^0$. Then,

$$E_{\gamma_0}\rho_{\theta\theta}^\dagger(W_i,\theta_0) = E_{\gamma_0}\rho_\theta^\dagger(W_i,\theta_0)\rho_\theta^\dagger(W_i,\theta_0)' = \phi_{1,0}\sum_{y,d=0,1}\tilde{M}_{yd,1}^\dagger + \phi_{0,0}\sum_{y,d=0,1}\tilde{M}_{yd,0}^\dagger. \tag{B.32}$$

For notational simplicity, let $c \equiv C_2\left(h_3(\pi_0),\zeta_{10};\pi_0\right)$ and $\tilde{c} \equiv C_2\left(h_2(\pi_0),\zeta_{10};\pi_0\right)$. Also let $d \equiv C_{\pi 2}\left(h_3(\pi_0),\zeta_{10};\pi_0\right) + C_{12}\left(h_3(\pi_0),\zeta_{10};\pi_0\right)h_{3,\pi}(\zeta_{10},\zeta_{30},\pi_0)$ and $\tilde{d} \equiv C_{\pi 2}\left(h_2(\pi_0),\zeta_{10};\pi_0\right) + C_{12}\left(h_2(\pi_0),\zeta_{10};\pi_0\right)h_{2,\pi}(\zeta_{10},\zeta_{20},\pi_0)$. Therefore,

$$M_{11,1}^\dagger = \begin{bmatrix} c^2 & 0 & 0 & c & cd \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ c & 0 & 0 & 1 & d \\ cd & 0 & 0 & d & d^2 \end{bmatrix}, \quad M_{01,1}^\dagger = \begin{bmatrix} (1-c)^2 & 1-c & 0 & c-1 & (c-1)d \\ 1-c & 1 & 0 & -1 & -d \\ 0 & 0 & 0 & 0 & 0 \\ c-1 & -1 & 0 & 1 & d \\ (c-1)d & -d & 0 & d & d^2 \end{bmatrix},$$

$$
M_{10,1}^\dagger = \begin{bmatrix} \tilde{c}^2 & 0 & -\tilde{c} & 0 & \tilde{c}\tilde{d} \\ 0 & 0 & 0 & 0 & 0 \\ -\tilde{c} & 0 & 1 & 0 & -\tilde{d} \\ 0 & 0 & 0 & 0 & 0 \\ \tilde{c}\tilde{d} & 0 & -\tilde{d} & 0 & \tilde{d}^2 \end{bmatrix}, \quad
M_{00,1}^\dagger = \begin{bmatrix} (1-\tilde{c})^2 & 1-\tilde{c} & 1-\tilde{c} & 0 & (\tilde{c}-1)\tilde{d} \\ 1-\tilde{c} & 1 & 1 & 0 & -\tilde{d} \\ 1-\tilde{c} & 1 & 1 & 0 & -\tilde{d} \\ 0 & 0 & 0 & 0 & 0 \\ (\tilde{c}-1)\tilde{d} & -\tilde{d} & -\tilde{d} & 0 & \tilde{d}^2 \end{bmatrix},
$$

$$
M_{11,0}^\dagger = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad
M_{01,0}^\dagger = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},
$$

$$
M_{10,0}^\dagger = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad
M_{00,0}^\dagger = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.
$$

By Lemma B.1, in analogy to the verification of S3(iv)(a), since $\sum_{y,d=0,1} \lambda_{\min}\left(\tilde{M}_{yd,0}^\dagger\right) = \lambda_{\min}(\tilde{M}_{00,1}^\dagger) = 0$, we consider the rest of the sum in (B.32) and apply (B.27). Let $a \equiv p_{11}^0/p_{01}^0$ and $b \equiv p_{11}^0/p_{10}^0$. Then, $\tilde{M}_{11,1}^\dagger + \tilde{M}_{01,1}^\dagger + \tilde{M}_{10,1}^\dagger = \left(M_{11,1}^\dagger + aM_{01,1}^\dagger + bM_{10,1}^\dagger\right)/p_{11}^0$ and

$$
\begin{aligned}
M^\dagger &\equiv M_{11,1}^\dagger + aM_{01,1}^\dagger + bM_{10,1}^\dagger \\
&= \begin{bmatrix}
a(1-c)^2 + b\tilde{c}^2 + c^2 & a(1-c) & -b\tilde{c} & -a(1-c)+c & a(c-1)d + b\tilde{c}\tilde{d} + cd \\
a(1-c) & a & 0 & -a & -ad \\
-b\tilde{c} & 0 & b & 0 & -b\tilde{d} \\
-a(1-c)+c & -a & 0 & a+1 & (a+1)d \\
a(c-1)d + b\tilde{c}\tilde{d} + cd & -ad & -b\tilde{d} & (a+1)d & (a+1)d^2 + b\tilde{d}^2
\end{bmatrix}.
\end{aligned}
$$

For the k-th leading principal minor $\left|M_k^\dagger\right|$ of $M^\dagger$,

$$
\left|M_1^\dagger\right| = a(1-c)^2 + b\tilde{c}^2 + c^2 > 0,
$$

$$
\left|M_2^\dagger\right| = ab\tilde{c}^2 + ac^2 > 0,
$$

$$\left|M_3^\dagger\right| = abc^2 > 0,$$

$$\left|M_4^\dagger\right| = a^2b(1-c)^2 + abc^2 > 0,$$

$$\left|M_5^\dagger\right| = \left|M^\dagger\right| = ab\left\{a^2(1+(1-c)^2)d^2 + b^2\tilde{c}^2\tilde{d}^2 + c^2(d^2+b\tilde{d}^2) + a\left(((1-c)^2+c^2)d^2 + bc^2\tilde{d}^2\right)\right\} > 0.$$

Therefore, $\tilde{M}_{01,1}^\dagger + \tilde{M}_{10,1}^\dagger + \tilde{M}_{11,1}^\dagger$ is positive definite and by (B.27), we can easily show that $\lambda_{\min}(E_{\gamma_0}\rho_{\theta\theta}^\dagger(W_i,\theta_0)) > 0$. ∎

Verification of S3(v): Recall

$$V^\dagger(\theta_1,\theta_2;\gamma_0) \equiv Cov_{\gamma_0}\left(\rho_\theta^\dagger(W_i,\theta_1),\rho_\theta^\dagger(W_i,\theta_2)\right).$$

But

$$Cov_{\gamma_0}\left(\rho_\theta^\dagger(W_i,\theta_0),\rho_\theta^\dagger(W_i,\theta_0)\right) = E_{\gamma_0}\rho_\theta^\dagger(W_i,\theta_0)\rho_\theta^\dagger(W_i,\theta_0)'$$
$$= E_{\gamma_0}\rho_{\theta\theta}^\dagger(W_i,\theta_0), \tag{B.33}$$

where the first equality is by $E_{\gamma_0}\rho_\theta^\dagger(W_i,\theta_0) = B^{-1}(\beta_0)E_{\gamma_0}\rho_\theta(w,\theta_0) = 0$ and the second equality is by the definition of $\rho_\theta^\dagger(W_i,\theta)$ and $\rho_{\theta\theta}^\dagger(W_i,\theta)$. Since $E_{\gamma_0}\rho_{\theta\theta}^\dagger(W_i,\theta_0)$ is positive definite from S3(iv)(b), we have the desired result. ∎

Define the $d_\psi \times d_\beta$ matrix-valued function

$$K(\theta;\gamma_0) \equiv \frac{\partial}{\partial\beta_0'}E_{\gamma_0}\rho_\psi(W_i,\theta) \tag{B.34}$$

with domain $\Theta_\delta \times \Gamma_0$, where $\Theta_\delta \equiv \{\theta \in \Theta : |\beta| < \delta\}$ and

$$\Gamma_0 \equiv \{\gamma_a = (a\beta,\zeta,\pi,\phi) \in \Gamma : \gamma = (\beta,\zeta,\pi,\phi) \in \Gamma \text{ with } |\beta| < \delta \text{ and } a \in [0,1]\}$$

for some $\delta > 0$.

Assumption S4: (i) $K(\theta;\gamma_0)$ exists $\forall(\theta,\gamma_0) \in \Theta_\delta \times \Gamma_0$.
(ii) $K(\theta;\gamma^*)$ is continuous in (θ,γ^*) at $(\theta,\gamma^*) = ((\psi_0,\pi),\gamma_0)$ uniformly over $\pi \in \Pi$ $\forall\gamma_0 \in \Gamma$ with $\beta_0 = 0$, where ψ_0 is a subvector of γ_0.

Verification of S4(i): Note that

$$K(\theta;\gamma_0) \equiv \frac{\partial}{\partial\beta_0}E_{\gamma_0}\rho_\psi(W_i,\theta)$$

$$= -\frac{\partial}{\partial \beta_0} \sum_{y,d,z=0,1} \frac{p_{yd,z}(\theta_0)\phi_{z,0}}{p_{yd,z}(\theta)} D_\psi p_{yd,z}(\theta)$$

$$= -\sum_{y,d,z=0,1} \frac{\partial p_{yd,z}(\theta_0)}{\partial \beta_0} \frac{\phi_{z,0}}{p_{yd,z}(\theta)} D_\psi p_{yd,z}(\theta),$$

where $\frac{\partial p_{yd,z}(\theta_0)}{\partial \beta_0}$ is the first element of $D_{\psi_0} p_{yd,z}(\theta_0)$ for all (y,d,z), whose expressions are above.
∎

Verification of S4(ii): For

$$K(\pi; \gamma_0) \equiv K(\psi_0, \pi; \gamma_0) = -\sum_{y,d,z=0,1} \frac{\partial p_{yd,z}(\theta_0)}{\partial \beta_0} \frac{\phi_{z,0}}{p_{yd,z}(\psi_0, \pi)} D_\psi p_{yd,z}(\psi_0, \pi),$$

let $a_{yd,z}(\pi, \theta_0, \phi_{1,0}) \equiv \frac{\partial p_{yd,z}(\theta_0)}{\partial \beta_0} \frac{\phi_{z,0}}{p_{yd,z}(\psi_0, \pi)} D_\psi p_{yd,z}(\psi_0, \pi)$ since $\phi_{0,0} = 1 - \phi_{1,0}$. Note that $a_{yd,z}(\pi, \theta_0, \phi_{1,0})$ is continuous in its arguments by Lemma B.1(iv). We can show that $a_{yd,z}(\pi, \theta_0, \phi_{1,0})$ is continuous uniformly in $\pi \in \Pi$ by applying the uniform convergence result in Lemma 9.2 of ACMLwp to $a_{yd,z}(\pi, \theta_n, \phi_{1,n}) - a_{yd,z}(\pi, \theta_0, \phi_{1,0})$, using (i) the pointwise convergence (i.e., pointwise continuity) above, (ii) $a_{yd,z}(\pi, \theta_0, \phi_{1,0})$'s differentiability in π with derivatives bounded over $\pi \in \Pi$ by Lemma B.1 and (iii) the compactness of Π (B1(iii) below). ∎

Next, we impose conditions on the parameter spaces Θ and Γ. Define $\Theta_\delta^* \equiv \{\theta \in \Theta^* : |\beta| < \delta\}$, where Θ^* is the true parameter space for θ. The "optimization parameter space" Θ satisfies:

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

The following general results are useful in verifying B1 and B2 below: for a continuous function f, (i) if a set A is compact, then $f(A)$ is compact and (ii) $f^{-1}(int(A)) \subset int(f^{-1}(A))$ for any set A in the range of f, where the latter is necessary and sufficient for continuity. Also note that by definition, for a proper function f, if B is compact, then $f^{-1}(B)$ is compact. Lastly, for a function f, if $A \subset B$ then $f(A) \subset f(B)$.

Verification of B1: TC3(ii) implies B1(i) since

$$int(\Theta) = int(\bar{h}^{-1}(\boldsymbol{\Theta})) \supset \bar{h}^{-1}(int(\boldsymbol{\Theta})) \supset \bar{h}^{-1}(\boldsymbol{\Theta}^*) = \Theta^*,$$

where the first $\supset$ is by the continuity of $\bar{h}$ and the second $\supset$ is by TC3(ii) and $\bar{h}^{-1}$ being a

function. For B1(ii), first note that given TC3(iii),

$$\bar{h}^{-1}(\boldsymbol{\Theta}) \supset \bar{h}^{-1}\left(\left\{\beta \in \mathbb{R}^{d_\beta} : \|\beta\| < \delta\right\} \times \boldsymbol{\mathcal{Z}}^0 \times \Pi\right) \supset \bar{h}^{-1}(\boldsymbol{\Theta}^*_\delta).$$

But $\bar{h}^{-1}(\boldsymbol{\Theta}) = \Theta$ and

$$\begin{aligned}
\bar{h}^{-1}(\boldsymbol{\Theta}^*_\delta) &= \left\{\theta \in \Theta^* : \bar{h}(\theta) \in \boldsymbol{\Theta}^*_\delta\right\} \\
&= \left\{\theta \in \Theta^* : \bar{h}(\theta) \in \boldsymbol{\Theta}^*, \left|\bar{h}_1(\theta)\right| < \delta\right\} \\
&= \left\{\theta \in \Theta^* : \theta \in \Theta^*, |\beta| < \delta\right\} \\
&= \Theta^*_\delta,
\end{aligned}$$

where the third equality is by $\bar{h}$ being a homeomorphism and $\bar{h}_1(\theta) = \beta$ being the first element of $\bar{h}$. Also, with $B_\delta \equiv \left\{\beta \in \mathbb{R}^{d_\beta} : |\beta| < \delta\right\}$,

$$\begin{aligned}
\bar{h}^{-1}(B_\delta \times \boldsymbol{\mathcal{Z}}^0 \times \Pi) &= \left\{\theta \in \Theta^* : \bar{h}(\theta) \in B_\delta \times \boldsymbol{\mathcal{Z}}^0 \times \Pi\right\} \\
&= B_\delta \times \left\{\mu \in \mathcal{M}^* : h(\mu) \in \boldsymbol{\mathcal{Z}}^0 \times \Pi\right\} \\
&= B_\delta \times h^{-1}(\boldsymbol{\mathcal{Z}}^0 \times \Pi) \\
&\equiv B_\delta \times \mathcal{Z}^0 \times \Pi,
\end{aligned}$$

where $\mathcal{M}^* = \left\{\mu \in \mathbb{R}^{d_\mu} : \theta = (\beta, \mu) \text{ for some } \theta \in \Theta^*\right\}$, the second equality holds since $\bar{h}(\theta) = (\beta, h(\mu))$ and the last equality holds by TC3(iv). Lastly, B1(iii) holds by TC3(i). ■

Assumption B2: (i) Γ is compact and $\Gamma = \{\gamma = (\theta, \phi) : \theta \in \Theta^*, \phi \in \Phi^*(\theta)\}$.
(ii) $\forall \delta > 0$, $\exists \gamma = (\beta, \zeta, \pi, \phi) \in \Gamma$ with $0 < \|\beta\| < \delta$.
(iii) $\forall \gamma = (\beta, \zeta, \pi, \phi) \in \Gamma$ with $0 < \|\beta\| < \delta$ for some $\delta > 0$, $\gamma_a = (a\beta, \zeta, \pi, \phi) \in \Gamma$ $\forall a \in [0, 1]$.

Verification of B2: Consider B2(i). Under TC4(i), define $\Phi^*(\theta)$ as $\Phi^*(\theta) \equiv \boldsymbol{\Phi}^*(\bar{h}(\theta))$. Since $\boldsymbol{\Gamma}$ is compact, $\boldsymbol{\Theta}^*$ and $\boldsymbol{\Phi}^*(\boldsymbol{\theta})$ are compact for $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*$. Thus, $\Theta^* = \bar{h}^{-1}(\boldsymbol{\Theta}^*)$ is compact by the properness of $\bar{h}$. Also given (B.4), we have

$$\Phi^*(\theta) \equiv \boldsymbol{\Phi}^*(\bar{h}(\theta)) = \boldsymbol{\Phi}^* = [0.01, 0.99],$$

which is compact. And therefore Γ is also compact. Next, TC4(ii) implies B2(ii). This is because, $\forall \delta > 0$, for $\boldsymbol{\gamma} = (\beta, \boldsymbol{\mu}, \phi)$ that satisfies TC4(ii), let γ in B2(ii) be $\gamma = (\beta, h^{-1}(\boldsymbol{\mu}), \phi)$, which is in Γ since $(\beta, \boldsymbol{\mu}) \in \boldsymbol{\Theta}^*$ implies $(\beta, h^{-1}(\boldsymbol{\mu})) = \bar{h}^{-1}(\beta, \boldsymbol{\mu}) \in \Theta^*$. To show that TC4(iii) implies B2(iii), note that for any $\gamma = (\beta, \zeta, \pi, \phi) \in \Gamma$ with $0 < |\beta| < \delta$ for some $\delta > 0$, $\boldsymbol{\gamma} = (\beta, h(\zeta, \pi), \phi) \in \boldsymbol{\Gamma}$. By TC4(iii), this implies that $\boldsymbol{\gamma}_a = (a\beta, h(\zeta, \pi), \phi) \in \boldsymbol{\Gamma}$ $\forall a \in [0, 1]$.

Therefore, $\gamma_a = (a\beta, h^{-1}(h(\zeta, \pi)), \phi) \in \Gamma$. ∎

Define a "weighted non-central chi-square" process $\{\xi(\pi; \gamma_0, b) : \pi \in \Pi\}$ by

$$\xi(\pi; \gamma_0, b) \equiv -\frac{1}{2} \left(G(\pi; \gamma_0) + K(\pi; \gamma_0)b \right)' H^{-1}(\pi; \gamma_0) \left(G(\pi; \gamma_0) + K(\pi; \gamma_0)b \right),$$

where $G(\pi; \gamma_0)$ is defined such that $G_n(\cdot) \Rightarrow G(\cdot; \gamma_0)$, where "$\Rightarrow$" denotes weak convergence, with

$$G_n(\pi) \equiv n^{-1/2} \sum_{i=1}^n \left(\rho_\psi(W_i; \psi_{0,n}, \pi) - E_{\gamma_n} \rho_\psi(W_i; \psi_{0,n}, \pi) \right)$$

and

$$H(\pi; \gamma_0) \equiv E_{\gamma_0} \rho_{\psi\psi}(W_i; \psi_0, \pi).$$

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

In Assumption C6, $\pi^*(\gamma_0, b)$ is random. The following is a primitive sufficient condition for Assumption C6 for the case where β is scalar. Let $\rho_\psi(w, \theta) \equiv (\rho_\beta(w, \theta)', \rho_\zeta(w, \theta)')'$. When $\beta = 0$, $\rho_\zeta(w, \theta)'$ does not depend on π by Assumption S2(ii) and is denoted by $\rho_\zeta(w, \psi)'$. For $\beta_0 = 0$, define

$$\rho_\psi^*(W_i, \psi_0, \pi_1, \pi_2)' \equiv (\rho_\beta(W_i, \psi_0, \pi_1)', \rho_\beta(W_i, \psi_0, \pi_2)', \rho_\zeta(W_i, \psi_0)')',$$

$$\Omega_G(\pi_1, \pi_2; \psi_0) \equiv Cov_{\gamma_0} \left(\rho_\psi^*(W_i, \psi_0, \pi_1, \pi_2)', \rho_\psi^*(W_i, \psi_0, \pi_1, \pi_2)' \right).$$

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

Note that Assumptions S1–S3 and C6† imply C6; see Lemma 3.1 of AC13.

Verification of C6†(ii): Noting that $D_\zeta p_{yd,z}(\psi_0, \pi)$ does not depend on π when $\beta_0 = 0$ so that we may denote it $D_\zeta p_{yd,z}(\psi_0)$, define

$$D_\psi p_{yd,z}^*(\psi_0, \pi_1, \pi_2) \equiv (D_\beta p_{yd,z}(\psi_0, \pi_1)', D_\beta p_{yd,z}(\psi_0, \pi_2)', D_\zeta p_{yd,z}(\psi_0)')'.$$

Then

$$\Omega_G(\pi_1, \pi_2; \psi_0) = E_{\gamma_0} \rho_\psi^*(W_i, \psi_0, \pi_1, \pi_2) \rho_\psi^*(W_i, \psi_0, \pi_1, \pi_2)'$$

$$= \sum_{y,d,z=0,1} \frac{\phi_{z,0}}{p_{yd}^0} D_\psi p_{yd,z}^*(\psi_0, \pi_1, \pi_2) D_\psi p_{yd,z}^*(\psi_0, \pi_1, \pi_2)',$$

where the second equality follows from (B.6) and $D_\psi p_{yd,z}^*(\psi_0, \pi_1, \pi_2)$ can be expressed as

$$D_\psi p_{11,0}^* = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad D_\psi p_{10,0}^* = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad D_\psi p_{01,0}^* = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ -1 \end{bmatrix}, \quad D_\theta p_{00,0}^* = \begin{bmatrix} 0 \\ 0 \\ -1 \\ -1 \\ 0 \end{bmatrix},$$

$$D_\psi p_{11,1}^* = \begin{bmatrix} C_2\left(h_3(\zeta_{10}, \zeta_{30}, \pi_1), \zeta_{10}; \pi_1\right) \\ C_2\left(h_3(\zeta_{10}, \zeta_{30}, \pi_2), \zeta_{10}; \pi_2\right) \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

$$D_\psi p_{10,1}^* = \begin{bmatrix} -C_2\left(h_2(\zeta_{10}, \zeta_{20}, \pi_1), \zeta_{10}; \pi_1\right) \\ -C_2\left(h_2(\zeta_{10}, \zeta_{20}, \pi_2), \zeta_{10}; \pi_2\right) \\ 0 \\ 1 \\ 0 \end{bmatrix},$$

and

$$D_\psi p_{01,1}^* = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} - D_\psi p_{11,1}^*, \quad D_\psi p_{00,1}^* = \begin{bmatrix} -1 \\ -1 \\ -1 \\ 0 \\ 0 \end{bmatrix} - D_\psi p_{10,1}^*,$$

using (B.23)-(B.26). The remaining arguments are similar to those used in the verification of S3(iv)(a): Let $M_{yd,z}^* \equiv D_\psi p_{yd,z}^*(\psi_0, \pi_1, \pi_2) \times D_\psi p_{yd,z}^*(\psi_0, \pi_1, \pi_2)'$ and $\tilde{M}_{yd,z}^* \equiv M_{yd,z}^*/p_{yd}^0$. Then,

$$\Omega_G(\pi_1, \pi_2; \psi_0) = \phi_{1,0} \sum_{y,d=1,1} \tilde{M}_{yd,1}^* + \phi_{0,0} \sum_{y,d=0,1} \tilde{M}_{yd,0}^*. \tag{B.35}$$

Let $c \equiv C_2\left(h_3(\zeta_{10}, \zeta_{30}, \pi_1), \zeta_{10}; \pi_1\right)$, $\tilde{c} \equiv C_2\left(h_3(\zeta_{10}, \zeta_{30}, \pi_2), \zeta_{10}; \pi_2\right)$, $d \equiv C_2\left(h_2(\zeta_{10}, \zeta_{20}, \pi_1), \zeta_{10}; \pi_1\right)$, and $\tilde{d} \equiv C_2\left(h_2(\zeta_{10}, \zeta_{20}, \pi_2), \zeta_{10}; \pi_2\right)$ for notational simplicity. Then,

$$
M_{11,1}^* = \begin{bmatrix} c^2 & c\tilde{c} & 0 & 0 & c \\ c\tilde{c} & \tilde{c}^2 & 0 & 0 & \tilde{c} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ c & \tilde{c} & 0 & 0 & 1 \end{bmatrix}, \quad
M_{01,1}^* = \begin{bmatrix} (1-c)^2 & (1-c)(1-\tilde{c}) & 1-c & 0 & -(1-c) \\ (1-c)(1-\tilde{c}) & (1-\tilde{c})^2 & 1-\tilde{c} & 0 & -(1-\tilde{c}) \\ 1-c & 1-\tilde{c} & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ -(1-c) & -(1-\tilde{c}) & -1 & 0 & 1 \end{bmatrix},
$$

$$
M_{10,1}^* = \begin{bmatrix} d^2 & d\tilde{d} & 0 & d & 0 \\ d\tilde{d} & \tilde{d}^2 & 0 & \tilde{d} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ d & \tilde{d} & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad
M_{00,1}^* = \begin{bmatrix} (1-d)^2 & (1-d)(1-\tilde{d}) & 1-d & 1-d & 0 \\ (1-d)(1-\tilde{d}) & (1-\tilde{d})^2 & 1-\tilde{d} & 1-\tilde{d} & 0 \\ 1-d & 1-\tilde{d} & 1 & 1 & 0 \\ 1-d & 1-\tilde{d} & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},
$$

$$
M_{11,0}^* = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad
M_{01,0}^* = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},
$$

$$
M_{10,0}^* = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad
M_{00,0}^* = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.
$$

By Lemma B.1 and similar arguments to those used to verify S3(iv)(a), since $\sum\limits_{y,d=0,1} \lambda_{\min}\left(\tilde{M}_{yd,0}^*\right) = \lambda_{\min}(\tilde{M}_{00,1}^*) = 0$, we consider the rest of the sum in (B.35) and apply (B.27). Let $a \equiv p_{01}^0/p_{10}^0$ and $b \equiv p_{01}^0/p_{11}^0$. Then, $\tilde{M}_{01,1}^* + \tilde{M}_{10,1}^* + \tilde{M}_{11,1}^* = \left(M_{01,1}^* + aM_{10,1}^* + bM_{11,1}^*\right)/p_{01}^0$, and

$$
M^* \equiv M_{01,1}^* + aM_{10,1}^* + bM_{11,1}^*
$$

$$
= \begin{bmatrix}
ad^2 + (1-c)^2 + bc^2 & ad\tilde{d} + (1-c)(1-\tilde{c}) + bc\tilde{c} & 1-c & ad & -(1-c)+bc \\
ad\tilde{d} + (1-c)(1-\tilde{c}) + bc\tilde{c} & a\tilde{d}^2 + (1-\tilde{c})^2 + b\tilde{c}^2 & 1-\tilde{c} & a\tilde{d} & -(1-\tilde{c})+b\tilde{c} \\
1-c & 1-\tilde{c} & 1 & 0 & -1 \\
ad & a\tilde{d} & 0 & a & 0 \\
-(1-c)+bc & -(1-\tilde{c})+b\tilde{c} & -1 & 0 & 1+b
\end{bmatrix}.
$$

For the k-th leading principal minor $|M_k^*|$ and determinant $|M^*|$ of M^*,

$$
|M_1^*| = ad^2 + (1-c)^2 > 0,
$$

$$
|M_2^*| = a \left\{ \tilde{d}(1-c) - d(1-\tilde{c}) \right\}^2 + b \left\{ c(1-\tilde{c}) - \tilde{c}^2(1-c) \right\}^2 + ab(\tilde{d}c - d\tilde{c})^2 > 0,
$$

$$
|M_3^*| = ab(\tilde{d}c - d\tilde{c})^2 + ab(\tilde{d}c + d\tilde{c})^2 + 4bc\tilde{c}(1-c)(1-\tilde{c}) > 0,
$$

$$
|M_4^*| = a \left\{ a\tilde{d}^2(1-c)^2 + (1-c)^2(1-\tilde{c})^2 \right\}
$$
$$
+ ab \left\{ (1-\tilde{c})^2 c^2 + bc^2\tilde{c}^2 + ad^2\tilde{c}^2 + (1-c)^2\tilde{c}^2 \right\} > 0,
$$

$$
|M^*| = ab \left[a(\tilde{d}c - d\tilde{c})^2 + \left\{ c(1-\tilde{c}) - \tilde{c}(1-c) \right\}^2 + a \left\{ (\tilde{d}(1-c) - d\tilde{c})^2 + (1-b)d^2\tilde{c}^2 \right\} \right.
$$
$$
\left. + a^2 d^2 \tilde{d}^2 + (1-c)^2(1-\tilde{c})^2 + b(1-\tilde{c})^2 c^2 + b^2 c^2\tilde{c}^2 + b\tilde{c}^2(1-c)^2 + 2a\tilde{d}^2 c(1-c) \right] > 0.
$$

Therefore, $\tilde{M}_{01,1}^* + \tilde{M}_{10,1}^* + \tilde{M}_{11,1}^*$ is positive definite and by (B.27), we can easily show that $\lambda_{\min}(\Omega_G(\pi_1, \pi_2; \psi_0)) > 0$. $\blacksquare$

Define a non-stochastic function $\{\eta(\pi; \gamma_0) : \pi \in \Pi\}$ by

$$
\eta(\pi; \gamma_0) \equiv -\frac{1}{2} K(\pi; \gamma_0)' H^{-1}(\pi; \gamma_0) K(\pi; \gamma_0).
$$

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

For $\beta_0 = 0$, by (B.15)–(B.16) we can write

$$
K(\pi; \gamma_0) = -\sum_{y,d,z=0,1} \frac{\phi_{z,0}}{p_{yd}^0} \frac{\partial p_{yd,z}(\theta_0)}{\partial \beta_0} D_\psi p_{yd,z}(\psi_0, \pi),
$$

$$
H(\pi; \gamma_0) = \sum_{y,d,z=0,1} \frac{\phi_{z,0}}{p_{yd}^0} D_\psi p_{yd,z}(\psi_0, \pi) D_\psi p_{yd,z}(\psi_0, \pi)'.
$$

Note that we can partition $H(\pi)$ and $K(\pi)$, suppressing γ_0, as

$$H(\pi) = \begin{bmatrix} H_{11}(\pi) & H_{12}(\pi) \\ H_{21}(\pi) & H_{22} \end{bmatrix} \begin{matrix} \}d_\beta \\ \}d_\zeta \end{matrix} \text{ and } K(\pi) = \begin{pmatrix} K_1(\pi) \\ K_2 \end{pmatrix} \begin{matrix} \}d_\beta \\ \}d_\zeta \end{matrix},$$

and note that $K(\pi_0) = [-H_{11}(\pi_0) : -H_{21}(\pi_0)']'$ by the expressions for $K(\pi; \gamma_0)$ and $H(\pi; \gamma_0)$.

Verification of C7: We first show that, for any $\pi \in \Pi$,

$$\eta(\pi) \geq \eta(\pi_0).$$

For matrices A and B, let $A \leq B$ denote $B - A$ being p.s.d. Then we can show that

$$K(\pi)'H^{-1}(\pi)K(\pi) \leq H_{11}(\pi_0) = K(\pi_0)'H^{-1}(\pi_0)K(\pi_0), \tag{B.36}$$

where the inequality is an application of the matrix Cauchy-Schwarz inequality (Proposition B.1 below) and the equality holds because $K(\pi_0) = [-H_{11}(\pi_0) : -H_{21}(\pi_0)']'$; see below for the proof. Lastly, the weak inequality in (B.36) holds as an equality if and only if $\rho_\beta(W_i, \psi_0, \pi_0)a + \rho_\psi(W_i, \psi_0, \pi)'b = 0$ with probability 1 for some $a \in \mathbb{R}$ and $b \in \mathbb{R}^{d_\psi}$ with $(a, b') \neq 0$. Let $D_\beta p^0_{yd,z} \equiv D_\beta p_{yd,z}(\psi_0, \pi_0)$ and $D_\psi p_{yd,z}(\pi) \equiv D_\psi p_{yd,z}(\psi_0, \pi)$ for simplicity. Then, when $\beta_0 = 0$

$$\rho_\beta(W_i, \psi_0, \pi_0)a + \rho_\psi(W_i, \psi_0, \pi)'b = \sum_{y,d,z=0,1} \frac{\mathbf{1}_{ydz}(W_i)}{p^0_{yd}} \left[D_\beta p^0_{yd,z}a + D_\psi p_{yd,z}(\pi)'b \right].$$

But, it is easy to see that a $(1 + d_\psi) \times 8$ matrix (suppressing π in $D_\psi p_{yd,z}(\pi)$ and letting $h_{3,0} \equiv h_3(\pi_0)$ and $h_{2,0} \equiv h_2(\pi_0)$)

$$\begin{bmatrix} D_\beta p^0_{11,1} & D_\beta p^0_{10,1} & D_\beta p^0_{01,1} & D_\beta p^0_{00,1} & D_\beta p^0_{11,0} & D_\beta p^0_{10,0} & D_\beta p^0_{01,0} & D_\beta p^0_{00,0} \\ D_\psi p_{11,1} & D_\psi p_{10,1} & D_\psi p_{01,1} & D_\psi p_{00,1} & D_\psi p_{11,0} & D_\psi p_{10,0} & D_\psi p_{01,0} & D_\psi p_{00,0} \end{bmatrix}$$

$$= \begin{bmatrix} C_2(h_{3,0}; \zeta_{10}, \pi_0) & -C_2(h_{2,0}, \zeta_{10}; \pi_0) & 1 - C_2(h_{3,0}, \zeta_{10}; \pi_0) & -1 + C_2(h_{2,0}, \zeta_{10}; \pi_0) & 0 & 0 & 0 & 0 \\ C_2(h_3; \zeta_{10}, \pi_0) & -C_2(h_2, \zeta_{10}; \pi_0) & 1 - C_2(h_3, \zeta_{10}; \pi_0) & -1 + C_2(h_2, \zeta_{10}; \pi_0) & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 \\ 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 \end{bmatrix}$$

has full row rank (i.e., rank of $1 + d_\psi$) except when $\pi = \pi_0$, since

$$C_2(h_3(\pi), \zeta_{10}; \pi) \neq C_2(h_3(\pi_0), \zeta_{10}; \pi_0)$$
$$C_2(h_2(\pi), \zeta_{10}; \pi) \neq C_2(h_2(\pi_0), \zeta_{10}; \pi_0)$$

75

for $\pi \neq \pi_0$. This can be shown by modifying the proof of Lemmas 3.1 and 4.1 of HV16 under Assumption TC2, which yields

$$\partial C_2 \left(h_3(\pi), \zeta_1; \pi\right) / \partial \pi = C_{\pi 2} \left(h_3(\pi), \zeta_1; \pi\right) + C_{12} \left(h_3(\pi), \zeta_1; \pi\right) h_{3,\pi}(\pi) < 0$$

and

$$C_{\pi 2} \left(h_2(\pi), \zeta_1; \pi\right) + C_{12} \left(h_2(\pi), \zeta_1; \pi\right) h_{2,\pi}(\pi) < 0.$$

In fact, h_2 or h_3 can be seen as u_1^* in Lemma 4.1 of HV16. Therefore, there is no (a, b') with $(a, b') \neq 0$ such that $D_\beta p_{yd,z}^0 a + D_\psi p_{yd,z}'(\pi) b = 0$ for all $(y, d, z) \in \{0, 1\}^3$, which implies that there is no (a, b') with $(a, b') \neq 0$ such that $\rho_\beta(W_i, \psi_0, \pi_0) a + \rho_\psi(W_i, \psi_0, \pi)' b = 0$ with probability 1. In other words, the equality holds uniquely at $\pi = \pi_0$ so that for any $\pi \neq \pi_0$, $\Pr[c'(\rho_\beta(W_i, \psi_0, \pi_0), \rho_\psi(W_i, \psi_0, \pi)')' = 0] < 1$ for all $c \in \mathbb{R}^{d_\beta + d_\psi}$ with $c \neq 0$ and thus the inequality in (B.36) is strict. $\blacksquare$

Proposition B.1. *Let $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$ be random vectors such that $E \|x\|^2 < \infty$, $E \|y\|^2 < \infty$, and Eyy' is nonsingular. Then*

$$\left(Exy'\right) \left(Eyy'\right)^{-1} \left(Eyx'\right) \leq Exx'.$$

For our verification proof, taking $x = \rho_\beta(W_i, \psi_0, \pi_0)$ and $y = \rho_\psi(W_i, \psi_0, \pi)$, we have

$$E_{\gamma_0} yy' = H(\pi),$$
$$E_{\gamma_0} xx' = H_{11}(\pi_0),$$
$$-E_{\gamma_0} xy' = -\left(E_{\gamma_0} yx'\right)' = K(\pi).$$

Proof of $H_{11}(\pi_0) = K(\pi_0)' H^{-1}(\pi_0) K(\pi_0)$**:** Define a 4×4 block-diagonalizing matrix

$$A(r) = \begin{bmatrix} 1 & -H_{12}(r) H_{22}^{-1} \\ 0_3 & I_3 \end{bmatrix}.$$

Then,

$$K(r_0)' H^{-1}(r_0) K(r_0) = K(r_0)' A(r) \left[A(r) H(r_0) A(r)\right]^{-1} A(r) K(r_0)$$

$$= (-1)^2 [H_{11}(r_0) : H_{21}(r_0)'] A(r) \left[A(r) H(r_0) A(r)\right]^{-1} A(r) \begin{bmatrix} H_{11}(r_0) \\ H_{21}(r_0) \end{bmatrix}$$

$$= [H_{11}(r_0) - H_{12}(r_0) H_{22}^{-1} H_{21}(r_0) : H_{21}(r_0)'] \begin{bmatrix} H_{11}^*(r_0)^{-1} & 0 \\ 0 & H_{22}^{-1} \end{bmatrix}$$

$$\times \left[\begin{array}{c} H_{11}(r_0) - H_{12}(r_0)H_{22}^{-1}H_{21}(r_0) \\ H_{21}(r_0) \end{array} \right]$$

$$= [1 : H_{21}(r_0)'H_{22}^{-1}] \left[\begin{array}{c} H_{11}(r_0) - H_{12}(r_0)H_{22}^{-1}H_{21}(r_0) \\ H_{21}(r_0) \end{array} \right]$$

$$= H_{11}(r_0),$$

where the second equality is due to the fact that $K(r_0) = [-H_{11}(r_0) : -H_{21}(r_0)']'$ and $H_{11}^*(r_0)$ is implicitly defined. We also use the symmetricity of $H(r)$ in this derivation. $\blacksquare$

Define the following quantities that arise in the asymptotic distribution of $\hat{\theta}_n$ and the test statistics we consider. Letting $S_\psi \equiv \left[I_{d_\psi} : 0_{d_\psi \times 1} \right]$ denote the $d_\psi \times d_\theta$ selector matrix that selects ψ out of θ:

$$\Omega(\pi_1, \pi_2; \gamma_0) \equiv S_\psi V^\dagger\left((\psi_0, \pi_1), (\psi_0, \pi_2); \gamma_0 \right) S_\psi',$$

$$J(\theta; \gamma_0) \equiv E_{\gamma_0} \rho_{\theta\theta}^\dagger(W_i; \theta),$$

$$V(\theta; \gamma_0) = V^\dagger(\theta, \theta; \gamma_0),$$

and

$$J(\gamma_0) \equiv J(\theta_0; \gamma_0),$$

$$V(\gamma_0) \equiv V(\theta_0; \gamma_0).$$

Note that

$$J(\gamma_0) = V(\gamma_0)$$

by (B.33). Define

$$\Sigma(\theta; \gamma_0) \equiv J^{-1}(\theta; \gamma_0)V(\theta; \gamma_0)J^{-1}(\theta; \gamma_0)$$

and

$$\Sigma(\pi; \gamma_0) \equiv \Sigma(\psi_0, \pi; \gamma_0).$$

Assumption V1: (i) $\widehat{J}_n = \widehat{J}_n(\widehat{\theta}_n)$ and $\widehat{V}_n = \widehat{V}_n(\widehat{\theta}_n)$ for some (stochastic) $d_\theta \times d_\theta$ matrix-valued functions $\widehat{J}_n(\theta)$ and $\widehat{V}_n(\theta)$ on Θ that satisfy $\sup_{\theta \in \Theta} \left\| \widehat{J}_n(\theta) - J(\theta; \gamma_0) \right\| \rightarrow_p 0$ and $\sup_{\theta \in \Theta} \left\| \widehat{V}_n(\theta) - V(\theta; \gamma_0) \right\| \rightarrow_p 0$ under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ with $\|b\| < \infty$.
(ii) $J(\theta; \gamma_0)$ and $V(\theta; \gamma_0)$ are continuous in θ on Θ $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$.
(iii) $\lambda_{\min}(\Sigma(\pi; \gamma_0)) > 0$ and $\lambda_{\max}(\Sigma(\pi; \gamma_0)) < \infty$ $\forall \pi \in \Pi$, $\forall \gamma_0 \in \Gamma$ with $\beta_0 = 0$.

Verification of V1(i): We define the following:

$$\widehat{J}_n(\theta) \equiv \frac{1}{n}\sum_{i=1}^n \rho_{\theta\theta}^\dagger(W_i, \theta) = \frac{1}{n}\sum_{i=1}^n \rho_\theta^\dagger(W_i, \theta)\rho_\theta^\dagger(W_i, \theta)'$$

$$= \frac{1}{n}\sum_{i=1}^n \sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i)\frac{D_\theta p_{yd,z}^\dagger(\theta)D_\theta p_{yd,z}^\dagger(\theta)'}{p_{yd,z}(\theta)^2},$$

where $D_\theta p_{yd,z}^\dagger(\theta_0)$ are defined above. Also,

$$\widehat{V}_n(\theta) \equiv \frac{1}{n}\sum_{i=1}^n \rho_\theta^\dagger(W_i, \theta)\rho_\theta^\dagger(W_i, \theta)'$$

$$= \frac{1}{n}\sum_{i=1}^n \sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i)\frac{D_\theta p_{yd,z}^\dagger(\theta)D_\theta p_{yd,z}^\dagger(\theta)'}{p_{yd,z}(\theta)^2} = \widehat{J}_n(\theta).$$

The rest of the proof follows from the uniform law of large numbers in Lemma 9.3 of ACMLwp with Assumptions S1 and S3 and Θ being compact. ∎

Verification of V1(ii): The continuity follows from the fact that the first and second derivatives of $p_{yd,z}(\theta)$ are continuous by Lemma B.1(vi). ∎

Verification of V1(iii): Note that

$$\Sigma(\pi; \gamma_0) = J^{-1}(\psi_0, \pi; \gamma_0)V(\psi_0, \pi; \gamma_0)J^{-1}(\psi_0, \pi; \gamma_0) = V^{-1}(\psi_0, \pi; \gamma_0)$$

since $V(\psi_0, \pi; \gamma_0) = J(\psi_0, \pi; \gamma_0)$. This is because

$$V(\psi_0, \pi; \gamma_0) = Cov_{\gamma_0}\left(\rho_\theta^\dagger(W_i, \psi_0, \pi), \rho_\theta^\dagger(W_i, \psi_0, \pi)\right) = E_{\gamma_0}\rho_\theta^\dagger(W_i; \psi_0, \pi)\rho_\theta^\dagger(W_i; \psi_0, \pi)'$$

$$= E_{\gamma_0}\rho_{\theta\theta}^\dagger(W_i; \theta)$$

where the last equality holds since $\rho_{\theta\theta}^\dagger(w, \theta) = \rho_\theta^\dagger(w, \theta)\rho_\theta^\dagger(w, \theta)'$, and the second-to-last equality holds since

$$E_{\gamma_0}\rho_\theta^\dagger(W_i, \psi_0, \pi) = -\sum_{y,d,z=0,1}\phi_{z,0}D_\theta p_{yd,z}^\dagger(\psi_0, \pi)$$

$$= -\sum_{y,d=0,1}\phi_{0,0}D_\theta p_{yd,0}^\dagger(\psi_0, \pi) - \sum_{y,d=0,1}\phi_{1,0}D_\theta p_{yd,1}^\dagger(\psi_0, \pi)$$

$$= 0.$$

Now, for the first part of V1(iii), note that since each element of the vectors in (B.28)–(B.31)

78

are bounded by TC2(iii) and B2(i), the elements of the matrix

$$V(\psi_0, \pi; \gamma_0) = E_{\gamma_0} \rho_\theta^\dagger(W_i; \psi_0, \pi) \rho_\theta^\dagger(W_i; \psi_0, \pi)' = \sum_{y,d,z=0,1} \frac{\phi_{z,0}}{p_{yd,0}} D_\theta p_{yd,z}^\dagger(\psi_0, \pi) D_\theta p_{yd,z}^\dagger(\psi_0, \pi)'$$

are bounded. For a $d \times d$ matrix A, $\sum_{i=1}^d |\lambda_i| \le \sum_{i,j=1}^d |A_{ij}|$ where the λ_i's are A's eigenvalues and the A_{ij}'s are A's elements. Therefore, $\lambda_{\max}(V(\psi_0, \pi; \gamma_0)) < \infty$. This implies that $\lambda_{\min}(V^{-1}(\psi_0, \pi; \gamma_0)) > 0$. By Lemma B.1, the proof of the second part is similar to the proofs of S3(iv)(b) and S3(v) and we can show that $\lambda_{\min}(V(\psi_0, \pi; \gamma_0)) > 0$, which implies that $\lambda_{\max}(V^{-1}(\psi_0, \pi; \gamma_0)) < \infty$. ∎

References

Altonji, J. G., Elder, T. E., Taber, C. R., 2005. An evaluation of instrumental variable strategies for estimating the effects of catholic schooling. Journal of Human resources 40 (4), 791–821. 7

Andrews, D. W. K., Cheng, X., 2012a. Estimation and inference with weak, semi-strong, and strong identificationith weak, semi-strong, and strong identification. Econometrica 80, 2153–2211. 1

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

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

Andrews, D. W. K., Cheng, X., 2013b. Supplemental appendices for 'Maximum likelihood estimation and uniform inference with sporadic identification failure', Journal of Econometrics Supplement. B

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

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

Andrews, D. W. K., Guggenberger, P., 2014. Identification- and singularity-robust inference for moment condition models, Cowles Foundation Discussion Paper No. 1978. 1, 2, 5.2, 7.3

Andrews, I., Mikusheva, A., 2016a. Conditional inference with a functional nuisance parameter. Econometrica 84, 1571–1612. 1, 2, 5.2, 22

Andrews, I., Mikusheva, A., 2016b. A geometric approach to weakly identified econometric models. Econometrica 84, 1249–1264. 1, 5.2

Antoine, B., Renault, E., 2009. Efficient GMM with nearly-weak instruments. Econometrics Journal 12, 135–171. 1, 4

Antoine, B., Renault, E., 2012. Efficient minimum distance estimation with multiple rates of convergence. Journal of Econometrics 170, 350–367. 1, 4

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

Dovonon, P., Renault, E., 2013. Testing for common conditionally heteroskedastic factors. Econometrica 81, 2561–2586. 1

Dunbar, G. R., Lewbel, A., Pendaku, K., 2013. Children's resources in collective households: Identification, estimation, and an application to child poverty in Malawi. American Economic Review 103, 438–471. 2.4

Escanciano, J. C., Zhu, L., 2013. Set inferences and sensitivity analysis in semiparametric conditionally identified models, Unpublished Manuscript, Indiana University, Department of Economics. 1, 3

Evans, W. N., Schwab, R. M., November 1995. Finishing high school and starting college: Do catholic schools make a difference? The Quarterly Journal of Economics 110 (4), 941–74. 7

Goldman, D., Bhattacharya, J., Mccaffrey, D., Duan, N., Leibowitz, A., Joyce, G., Morton, S., 2001. Effect of Insurance on Mortality in an HIV-Positive Population in Care. Journal of the American Statistical Association 96 (455). 7

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

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

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., 2017. Identification in a generalization of bivariate probit models with dummy endogenous regressors. Journal of Econometrics 199, 63–73. 2.3, 7, B

Heckman, J. J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161. 2.1

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

Jones, D. R., 2001. A taxonomy of globabl optimization methods based on response surfaces. Journal of Globabl Optimization 21, 345–383. 6.2

Jones, D. R., Schonlau, M., Welch, W. J., 1998. Efficient globabl optimization of expensive black-box functions. Journal of Globabl Optimization 13, 455–492. 6.2

Kleibergen, F., 2002. Pivotal statistics for testing structural parameters in instromental variables regression. Econometrica 70, 1781–1803. 1, 7

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

Komunjer, I., Ng, S., 2011. Dynamic identification of dynamic stochastic general equilibrium models. Econometrica 79, 1995–2032. 3.4

Lee, L. F., Chesher, A., 1986. Specification testing when score statistics are identically zero. Journal of Econometrics 31, 121–149. 1

Lindelöf, E., 1894. Sur l'application des méthodes d'approximations successives à l'étude des intégrales réelles des équations différentielles ordinaires. Journal de Mathématiques Pures et Appliquées, 117–128. A

Lochner, L., Moretti, E., 2004. The effect of education on crime: evidence from prison inmates, arrests, and self-reports. American Economic Review 94, 155–189. 7, 8

McCloskey, A., 2017. Bonferroni-based size-correction for nonstandard testing problems, forthcoming in Journal of Econometrics. 1, 6.2, 6.2, A

Moreira, M. J., 2003. A conditional likelihood ratio test for structural models. Econometrica 71, 1027–1048. 1, 7

Phillips, P. C. B., 1989. Partially identified econometric models. Econometric Theory 5, 181–240. 1, 3.3, 4

Phillips, P. C. B., 2016. Inference in near-singular regression. In: G. Gonzalez-Rivera, R. C. H., Lee, T.-H. (Eds.), Essays in Honor of Aman Ullah. Vol. 36 of Advances in Econometrics. Emerald Group Publishing Limited, pp. 461–486. 1, 3.3, 4

Picard, E., 1893. Sur l'application des méthodes d'approximations successives à l'étude de certaines équations différentielles ordinaires. Journal de Mathématiques Pures et Appliquées, 217–272. A

Qu, Z., Tkachenko, D., 2012. Identification and frequency domain quasi-maximum likelihood estimation of linearized dynamic stochastic general equilibrium models. Quantitative Economics 3, 95–132. 1, 3, 3.4

Rhine, S. L., Greene, W. H., Toussaint-Comeau, M., 2006. The importance of check-cashing businesses to the unbanked: Racial/ethnic differences. Review of Economics and Statistics 88 (1), 146–157. 7

Rothenberg, T. J., 1971. Identification in parametric models. Econometrica 39, 577–591. 1, 3

Sargan, J. D., 1983. Identification and lack of identification. Econometrica 51, 1605–1633. 1, 1, 4

Sims, C. A., Stock, J. H., Watson, M. W., 1990. Inference in linear time series models with some unit roots. Econometrica 58, 113–144. 4

Staiger, D., Stock, J. H., 1997. Instrumental variables regression with weak instruments. Econometrica 65, 557–586. 1, 7

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

Tommasi, D., Wolf, A., 2016. Overcoming weak identification in the estimation of household resource shares, Unpublished Manuscript, ECARES. 2.4

Weyl, H., 1912. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). Mathematische Annalen 71 (4), 441–479. B.3

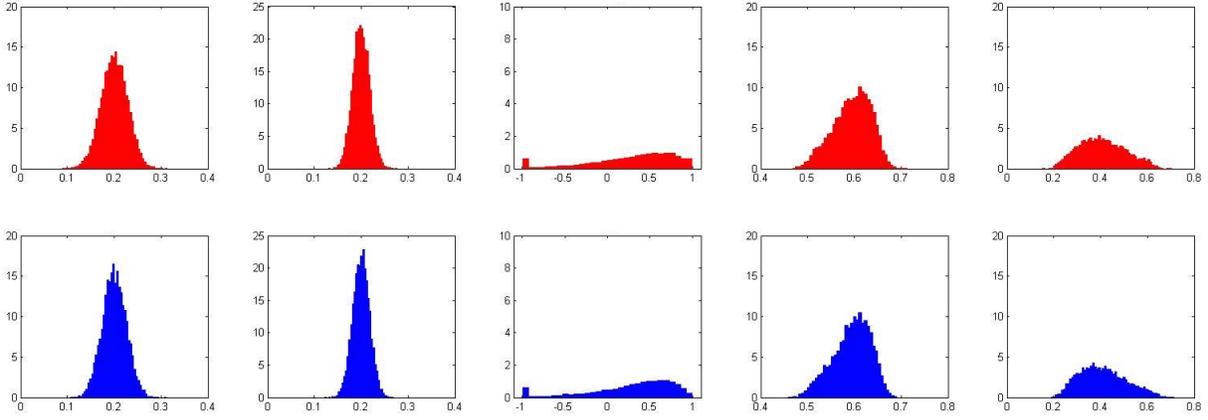Figure 1: Threshold Crossing Model Parameter Estimator Densities when $b = 0$
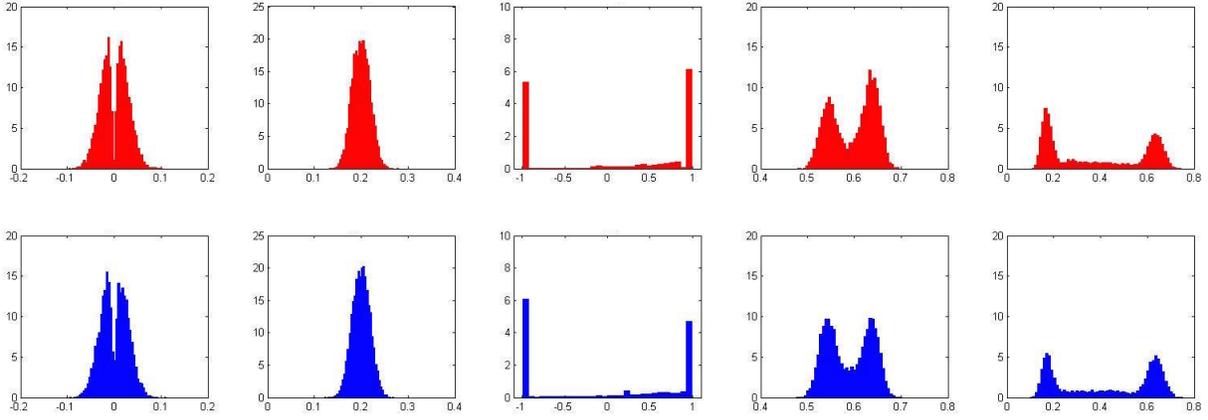
Asymptotic (blue) and finite-sample (red, $n = 1000$) densities of the estimators of β, $\boldsymbol{\zeta}$, $\boldsymbol{\pi}_3$, $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ (left-to-right) in the Threshold-Crossing model when $\boldsymbol{\zeta} = 0.2$ and $\boldsymbol{\pi} = (0.6, 0.4, 0.4)$.

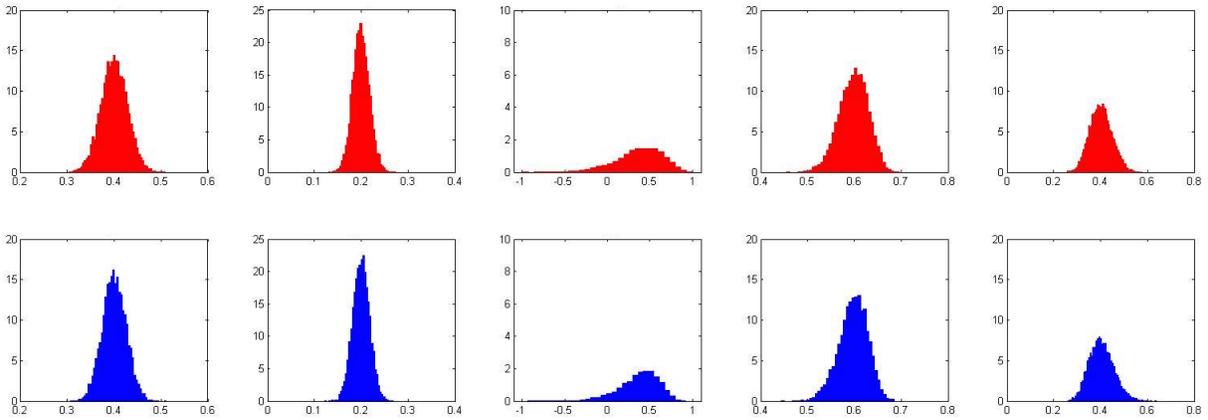Figure 2: Threshold Crossing Model Parameter Estimator Densities when $b = \sqrt{n}0.1$
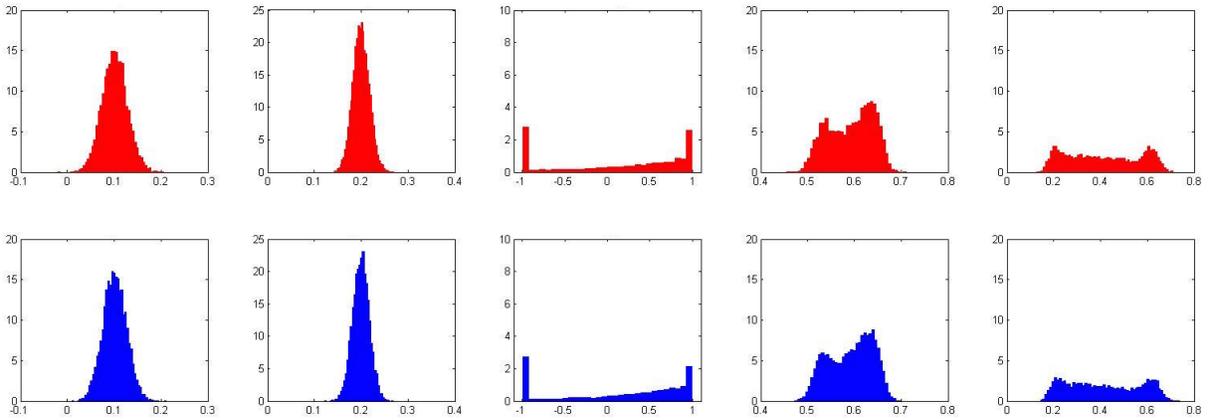
Asymptotic (blue) and finite-sample (red, $n = 1000$) densities of the estimators of β, $\boldsymbol{\zeta}$, $\boldsymbol{\pi}_3$, $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ (left-to-right) in the Threshold-Crossing model when $\boldsymbol{\zeta} = 0.2$ and $\boldsymbol{\pi} = (0.6, 0.4, 0.4)$.

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

Asymptotic (blue) and finite-sample (red, $n = 1000$) densities of the estimators of β, ζ, π_3, π_1 and π_2 (left-to-right) in the Threshold-Crossing model when $\zeta = 0.2$ and $\pi = (0.6, 0.4, 0.4)$.

Figure 4: Threshold Crossing Model Parameter Estimator Densities when $b = \sqrt{n}0.4$
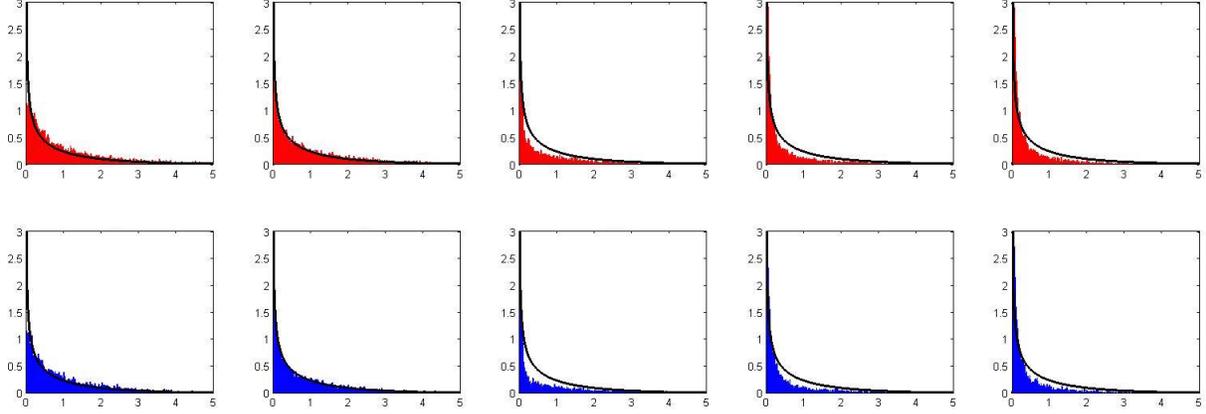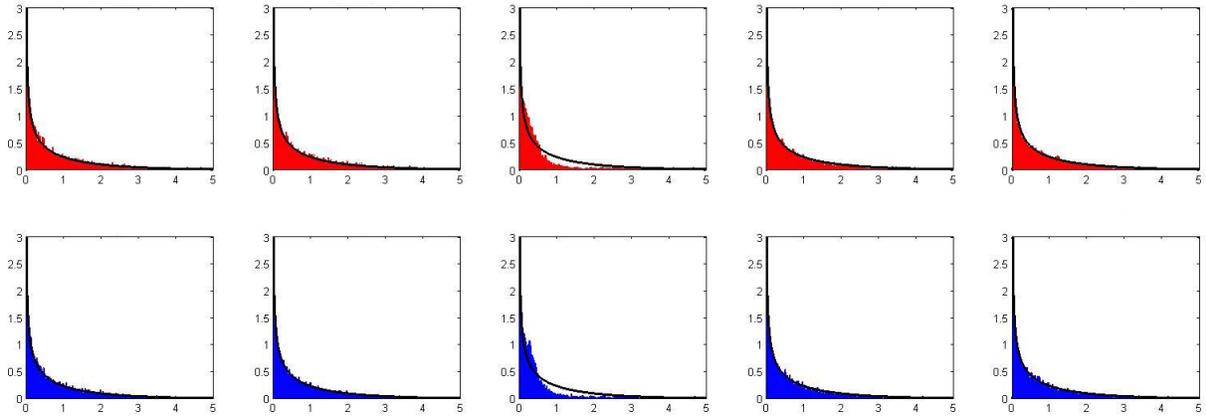
Asymptotic (blue) and finite-sample (red, $n = 1000$) densities of the estimators of β, ζ, π_3, π_1 and π_2 (left-to-right) in the Threshold-Crossing model when $\zeta = 0.2$ and $\pi = (0.6, 0.4, 0.4)$.

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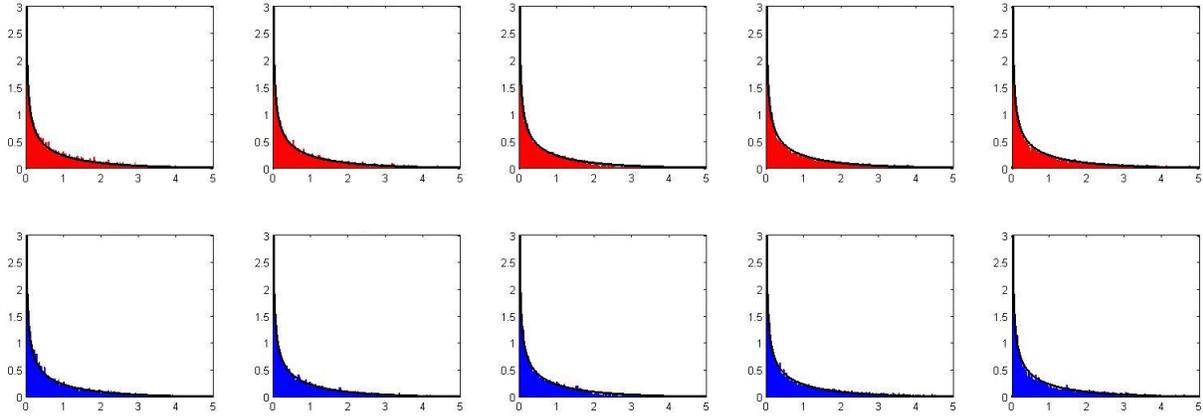

Asymptotic (blue) and finite-sample (red, $n = 1000$) densities of the Wald statistic for the parameters β, ζ, π_3, π_1 and π_2 (left-to-right) in the Threshold-Crossing model when $\zeta = 0.2$ and $\pi = (0.6, 0.4, 0.4)$, with a χ_1^2 density overlay (black line).

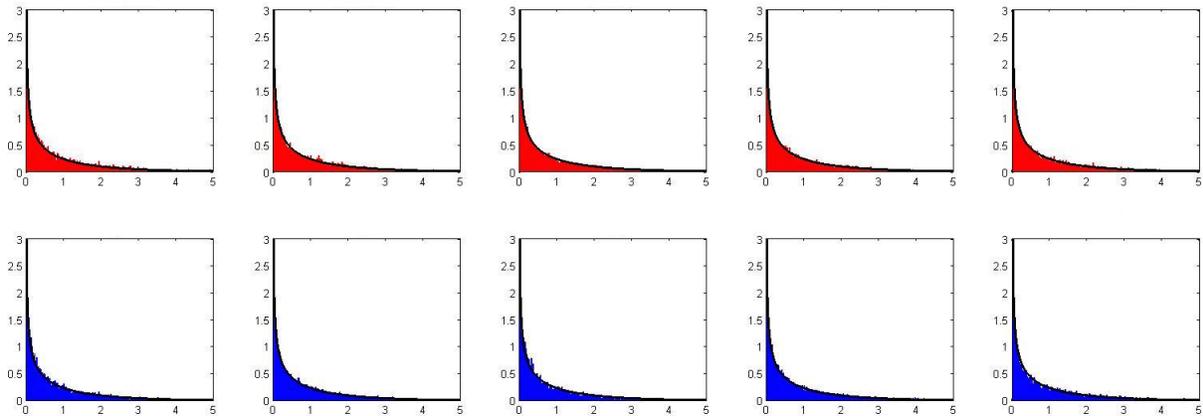Figure 6: Wald Statistic Densities for the Threshold Crossing Model when $b = \sqrt{n}0.1$

Asymptotic (blue) and finite-sample (red, $n = 1000$) densities of the Wald statistic for the parameters β, ζ, π_3, π_1 and π_2 (left-to-right) in the Threshold-Crossing model when $\zeta = 0.2$ and $\pi = (0.6, 0.4, 0.4)$, with a χ_1^2 density overlay (black line).

86

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

Asymptotic (blue) and finite-sample (red, $n = 1000$) densities of the Wald statistic for the parameters β, $\boldsymbol{\zeta}$, $\boldsymbol{\pi}_3$, $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ (left-to-right) in the Threshold-Crossing model when $\boldsymbol{\zeta} = 0.2$ and $\boldsymbol{\pi} = (0.6, 0.4, 0.4)$, with a χ^2_1 density overlay (black line).

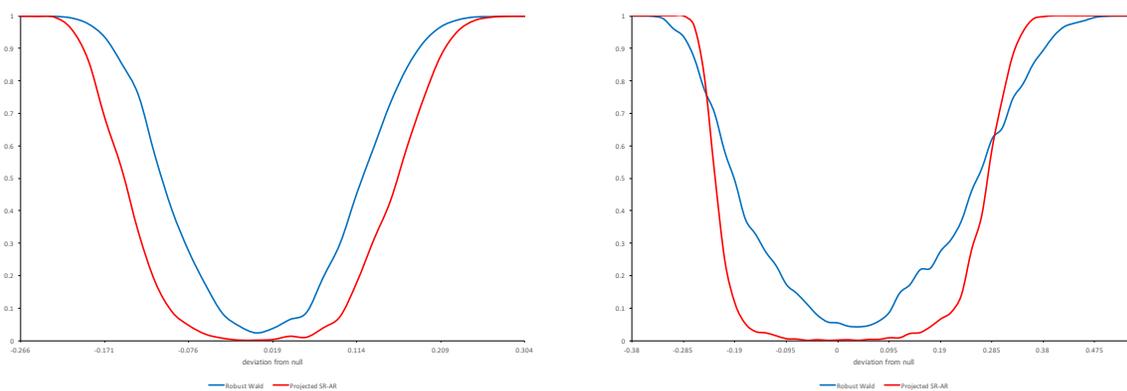Figure 8: Wald Statistic Densities for the Threshold Crossing Model when $b = \sqrt{n}0.4$

Asymptotic (blue) and finite-sample (red, $n = 1000$) densities of the Wald statistic for the parameters β, $\boldsymbol{\zeta}$, $\boldsymbol{\pi}_3$, $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ (left-to-right) in the Threshold-Crossing model when $\boldsymbol{\zeta} = 0.2$ and $\boldsymbol{\pi} = (0.6, 0.4, 0.4)$, with a χ^2_1 density overlay (black line).

Figure 9: Power Curves for Testing $\boldsymbol{\pi}_2$ in the Threshold Crossing Model

Robust Wald (blue) and projected SR-AR (red) power for testing $\boldsymbol{\pi}_2 = 0.4$ in the Threshold-Crossing model with $n = 1000$, when $\beta = 0.4$ (left - corresponding to strong identification) and $\beta = 0.2$ (right - corresponding to weak identification), $\boldsymbol{\zeta} = 0.2$, $\boldsymbol{\pi}_1 = 0.6$, $\boldsymbol{\pi}_3 = 0.4$ and $(\boldsymbol{\pi}_2 - 0.4)$ varies across the horizontal axes.