Quantitative Economics 10 (2019), 1019–1068 1759-7331/20191019

Estimation and inference with a (nearly) singular Jacobian

Sukjin Han
Department of Economics, University of Texas at Austin

Adam McCloskey
Department of Economics, University of Colorado at Boulder

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 underidentified and the identification status of individual parameters is not generally straightforward to characterize. We provide a systematic reparameterization procedure that leads to a reparametrized 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 nonidentification 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.

Keywords. Reparameterization, deficient rank Jacobian, asymptotic size, uniform inference, subvector inference, extremum estimators, identification, nonlinear models, Wald test, weak identification, underidentification.

JEL classification. C12, C15.

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

Sukjin Han: sukjin.han@austin.utexas.edu
Adam McCloskey: adam.mccloskey@colorado.edu

The authors are grateful to Donald Andrews, Isaiah Andrews, Xiaohong Chen, Xu Cheng, Gregory Cox, Áureo de Paula, Stephen Donald, Bruce Hansen, Bo Honoré, Tassos Magdalinos, Peter Phillips, Eric Renault, Jesse Shapiro, James Stock, Yixiao Sun, Elie Tamer, Edward Vytlacil, and four anonymous referees for helpful comments. This paper is developed from earlier work by Han (2010). The second author gratefully acknowledges support from the NSF under grant SES-1357607.

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 nonidentification scenario is sometimes referred to as "underidentification" or "partial identification." The nonidentification 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 nonidentification 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 underidentified/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 (global) identification failure in this paper through the Jacobian matrix of the model restrictions: the Jacobian matrix of the model restrictions has deficient column rank in a (typically linear) subspace of the entire parameter space.[1] We examine models for which a vector of parameters governs the identification status of the model, with identification failure occurring when this vector of parameters is equal to a specific value. 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 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 nonidentification, weak identification, and strong identification. We find that the asymptotic distributions derived under certain sequences of data-generating processes (DGPs) indexed by the sample size provide much better approximations to the finite sample distributions of these objects than those derived under the standard limit theory that assumes strong identification. Third, we use the limit theory derived under weak identification DGP sequences to construct data-dependent critical values (CVs) for Wald statistics that yield (uniformly) correct asymptotic size and good power properties. Importantly, our robust inference procedures allow one to directly conduct hypothesis tests for low-dimensional functions of the model parameters, including one-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)),

[1] See Rothenberg (1971) for a discussion of local versus global identification.

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) focused on a zero rank Jacobian as the point of identification failure in moment condition models. In this case, the identified set becomes the entire parameter space at points of identification failure. The recent works of Andrews and Cheng (2012, 2013, 2014) implicitly focus on models for which the Jacobian of the model restrictions has columns of zeros at points of identification failure. For these types of models, some parameters become "completely" unidentified (those corresponding to the zero columns) while others remain strongly identified. In this paper, we do not 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 (Forthcoming) and Andrews and Mikusheva (2016a) in the context of moment condition models. Andrews and Mikusheva (2016b) provided methods to directly perform uniformly valid subvector inference while Andrews and Guggenberger (Forthcoming) and Andrews and Mikusheva (2016a) do not.[2] Unlike these papers, but like Andrews and Cheng (2012, 2013, 2014), 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 (2012, 2013, 2014), we derive nonstandard limit theory for standard estimators and test statistics. This nonstandard limit theory

[2] Andrews and Mikusheva (2016a) provided 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 (Forthcoming) 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. We refer the interested reader to Remark 5.2 for further comparisons with the methods in these papers as well as Andrews and Mikusheva (2016b).

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 underidentification in different contexts. For example, Sargan (1983) studied regression models that are nonlinear in parameters and first-order locally underidentified. Phillips (1989) and Choi and Phillips (1992) studied underidentified simultaneous equations models and spurious time series regressions. In a rather different context, Lee and Chesher (1986) also made use of a reparameterization for a type of identification problem. Arellano, Hansen, and Sentana (2012) proposed a way to test for underidentification in a generalized method of moments (GMM) context. Qu and Tkachenko (2012) studied underidentification in the context of DSGE models. Escanciano and Zhu (2013) studied underidentification in a class of semiparametric models.[3] Dovonon and Renault (2013) uncovered an interesting result that, when testing for common sources of conditional heteroskedasticity in a vector of time series, there is a loss of first-order identification under the null hypothesis while the model remains second-order identified. Although all of these papers study underidentification of various forms, none of them deal with the empirically relevant potential for near or local to underidentification, 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 (2012) (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 reparametrized 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 reparametrized parameter estimates. Though beyond the scope of the current paper, our reparameterization procedure may similarly be useful as a step toward using the general limit theory of Cox (2017) for some problems. This latter paper studies other, more complicated forms of weak identification not covered by AC12.

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

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

[4]This corresponds to Assumption A of AC12.

reparametrized 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), Phillips (1989) and Choi and Phillips (1992) used similar rotations to derive limit theory for estimators under identification failure; Antoine and Renault (2009, 2012) used similar rotations to derive limit theory for estimators under "nearly-weak" identification;[5] Andrews and Cheng (2014) (AC14 hereafter) used similar rotations to find the asymptotic distributions of Wald statistics under weak and nearly-strong identification; and recently Phillips (2016) used 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 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 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 theoretical results of this paper are based upon widely applicable high-level assumptions. We verify these assumptions for the triangular threshold crossing model by imposing lower-level conditions in Online Appendix B in the Online Supplemental Material (Han and McCloskey (2019)).

The paper is organized as follows. In the next section, we introduce the general class of models subject to underidentification 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

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

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 and verification of assumptions for the threshold crossing model are contained in the in the Online Appendix (Supplemental Material, Han and McCloskey (2019)), while figures are collected at the end of the document. In addition, some of the assumptions and expressions from AC12 that also appear in this paper are collected for the reader's convenience in the Appendix at the end of the paper.

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 $f(a)$. Finally, we write "wp1" as shorthand for "with probability one."

2. General class of models

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

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

when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. The "model restriction" function describing this relationship $\boldsymbol{g}$ may depend on the true underlying parameter $\boldsymbol{\gamma}^*$ that contains $\boldsymbol{\theta}^*$, that is, the true underlying DGP. The parameter $\boldsymbol{\gamma}^*$ can be infinite-dimensional so that, for example, moment conditions may constitute (some of) the model restrictions. 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 $\boldsymbol{\theta} = \boldsymbol{\theta}^*$; see Example 2.3 below.

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 underidentified; and when β is "close" to $\bar{\beta}$ relative to sampling variability, $\boldsymbol{\mu}$ is local-to-underidentified. 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] We are considering models that become globally underidentified in a (typically linear) subspace 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).

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 $\beta = 0$, $\operatorname{rank}(\partial \boldsymbol{g}^*(\boldsymbol{\theta})/\partial \boldsymbol{\pi}') < d_{\boldsymbol{\pi}}$ 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

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

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).

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

where $X_i \equiv (1, X_{1i}')'$ is $k \times 1$, $Z_i \equiv (1, Z_{1i}')'$ is $l \times 1$, the variables (X_i, Z_i) are independent of the errors (ε_i, ν_i) and $(X_i, Z_i, \varepsilon_i, \nu_i)$ are i.i.d. Note that Z_i may include (components of) X_i. We observe $W_i = (D_i Y_i, D_i, X_i, Z_i)$ 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 $\boldsymbol{\theta} = \boldsymbol{\theta}^*$,

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

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

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

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

with $w = (dy, d, x, z)$ being a realization of W_i 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).

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

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).

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

where $Z_i \in \{0, 1\}$. We observe an i.i.d. sample of $W_i = (Y_i, D_i, Z_i)$ and assume that the instrument Z_i is independent of (ε_i, ν_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 $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, $\xi^* - \mathfrak{g}(\boldsymbol{\theta}) = 0$, where

[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.

$\xi = (p_{11,0}, p_{11,1}, p_{10,0}, p_{10,1}, p_{01,0}, p_{01,1})'$ with $p_{yd,z} \equiv \Pr_{\boldsymbol{\gamma}}[Y = y, D = d | Z = z]$ and

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

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

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

Example 2.4 (Engel curve models for household share). Tommasi and Wolf (2018) discussed Engel curve estimation for the private assignable good in the Dunbar, Lewbel, and Pendaku (2013) collective model for household expenditure shares when using the PIGLOG utility function. See equation (2) of Tommasi and Wolf (2018) 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 $0 = \boldsymbol{g}^*(\boldsymbol{\theta}) = E_{\boldsymbol{\gamma}^*}\boldsymbol{\varphi}(W_i, \boldsymbol{\theta})$ when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, where $\boldsymbol{\theta} = (\beta, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3)$ and the moment function is

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

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

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

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

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

with q_i, ∂q_i, $\Phi_i \equiv \Phi(\boldsymbol{\zeta} + Z_{1i}'\beta)$ and $\phi_i \equiv \phi(\boldsymbol{\zeta} + Z_{1i}'\beta)$ being the terms that depend on $(\beta, \boldsymbol{\zeta})$. Note that $d_{\boldsymbol{\zeta}} < \operatorname{rank}(J^*(\boldsymbol{\theta})) < d_{\boldsymbol{\mu}}$ when $\beta = 0$, since q_i becomes a constant and $X_i = (1, X_{1i}')'$. This type of behavior for the Jacobian matrix, which is common to many models, motivates the following assumption.

Assumption ID. *When* $\beta = 0$, $\operatorname{rank}(J^*(\boldsymbol{\theta})) \equiv r < d_{\boldsymbol{\mu}}$ *for all* $\boldsymbol{\theta} = (0, \boldsymbol{\mu}) \in \boldsymbol{\Theta}$.

In general, a rank-deficient Jacobian with nonzero 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 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 underidentified 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 models that are globally underidentified across their entire parameter space (in contrast to those that lose identification in the subspace 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 $\boldsymbol{\Theta}$:

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

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, $\bar{\boldsymbol{g}}_n(\boldsymbol{\theta}) = \hat{\xi}_n - \mathfrak{g}(\boldsymbol{\theta})$, where $\hat{\xi}_n$ is a sample analog of ξ^*, in analogy to (2.2). For GMM, $\bar{\boldsymbol{g}}_n(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \varphi(W_i, \boldsymbol{\theta})$.

ASSUMPTION CF. *$Q_n(\boldsymbol{\theta})$ can be written as*

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

*for some random function $\Psi_n(\cdot)$ that is differentiable wp*1.

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 certain types of likelihoods as in Example 2.3 involve $\boldsymbol{g}^*(\boldsymbol{\theta}) = \xi^* - \mathfrak{g}(\boldsymbol{\theta})$ by (2.2). For a GMM/MD criterion function, $\Psi_n(\bar{\boldsymbol{g}}_n(\boldsymbol{\theta})) = \|W_n\bar{\boldsymbol{g}}_n(\boldsymbol{\theta})\|^2$ where W_n is a (possibly random) weight matrix.[8] Note also that our framework includes general ML estimation with concave likelihoods, since it is numerically equivalent to GMM estimation that uses the score equations as moments.

ASSUMPTION REG1. *$\bar{\boldsymbol{g}}_n : \boldsymbol{\Theta} \to \mathbb{R}^{d_g}$ is continuously differentiable in $\boldsymbol{\theta}$ wp*1.

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 does not depend on the true DGP.

ASSUMPTION JAC. *When $\beta = 0$, the null space of $J^*(\boldsymbol{\theta})$ is a subspace of the null space of $\partial\bar{\boldsymbol{g}}_n(\boldsymbol{\theta})/\partial\boldsymbol{\mu}'$ wp*1 *for all $n \geq 1$. The null space of $J^*(\boldsymbol{\theta})$ does not depend upon the true DGP γ^*.*

We allow the null space of the sample Jacobian to be larger than that of the population Jacobian to accommodate the possibility that particular realizations of the random variables entering the sample Jacobian can induce additional rank reduction. For example, this allows the random variables X_{1i} in Example 2.1 above to equal zero for all $i = 1, \ldots, n$. Examples 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 DGP-dependent, but we have not found an application for which such an extension would be useful.

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 Jacobian $J^*(\boldsymbol{\theta})$ under identification failure, that is, $d_\pi \equiv d_{\boldsymbol{\mu}} - r$ (this will later denote the dimension of a new parameter π). Let the parameter space for $\boldsymbol{\mu}$ be denoted as

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

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

PROCEDURE 3.1. For a given $J^*(\boldsymbol{\theta})$ that satisfies Assumptions ID and Jac, 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:

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

1. Find a full rank $d_{\boldsymbol{\mu}} \times d_{\boldsymbol{\mu}}$ matrix M that performs elementary column operations[9] such that when $\beta = 0$,

$$J^*(\boldsymbol{\theta})M(\boldsymbol{\mu}) = \left[\boldsymbol{G}(\boldsymbol{\mu}) : 0_{d_g \times d_\pi}\right] \tag{3.1}$$

for all $\boldsymbol{\mu} \in \boldsymbol{\mathcal{M}}$, where $\boldsymbol{G}_n(\boldsymbol{\mu})$ is some $d_g \times r$ matrix.

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

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

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

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

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 decomposition can be used to compute the matrix $M(\boldsymbol{\mu})$ with conventional software since the right singular vectors of $J^*(\boldsymbol{\theta})$ that correspond to its zero singular values span its null space and its left singular vectors that correspond to its nonzero singular values span its column space.[10] 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 $(\beta, \boldsymbol{\mu}) = (\beta, h(\mu))$. Throughout the paper, we use boldface font for the original parameters and standard font for the transformed parameters; once all of the relevant parameters are introduced, we summarize the notation in Table 1. 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.

TABLE 1. Summary of notation.

$(\beta, \boldsymbol{\mu}) = (\beta, h(\mu))$	Determines ID	ID subject to failure	ID not subject to failure
Original parameters $\boldsymbol{\theta} \equiv (\beta, \boldsymbol{\mu}) \equiv (\beta, \boldsymbol{\zeta}, \boldsymbol{\pi})$	β	$\boldsymbol{\mu}, \boldsymbol{\pi}$	$\beta, \boldsymbol{\zeta}$
Transformed parameters $\theta \equiv (\beta, \mu) \equiv (\beta, \zeta, \pi)$	β	μ, π	β, ζ

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

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

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 the case for which $F_{\varepsilon\nu}(\varepsilon, \nu; \boldsymbol{\pi}^2)$ is a bivariate standard normal distribution, the Jacobian for this model with respect to $\boldsymbol{\mu}$ is

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

when $\beta = 0$, where $L_i(\beta, \boldsymbol{\zeta})$ is defined in (2.8). Suppose that $E_{\gamma^*}[D_i X_i X_i']$ has full rank. Then 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:

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

This yields

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

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 population and sample model restrictions and the criterion functions of the new parameter θ as

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

and

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

The new Jacobian $\partial g^*(\theta)/\partial \mu' = (\partial \boldsymbol{g}^*(\boldsymbol{\theta})/\partial \boldsymbol{\mu}')(\partial h(\mu)/\partial \mu')$ has the same reduced rank $r < d_\mu = d_{\boldsymbol{\mu}}$ as the original Jacobian $J^*(\boldsymbol{\theta}) = \partial \boldsymbol{g}^*(\boldsymbol{\theta})/\partial \boldsymbol{\mu}'$ 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

$\partial g^*(\theta)/\partial \mu'$ being equal to zero. Using this result, in conjunction with Assumption Jac, the reparametrized criterion function $Q_n(\theta)$ satisfies a property that is instrumental to deriving the limit theory detailed below.

THEOREM 3.1. *Under Assumptions* ID, CF, Reg1, *and* Jac, $Q_n(\theta)$ *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 *reparametrized 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 reparametrized estimation problem and is defined in terms of the original optimization parameter space $\boldsymbol{\Theta}$ as follows:

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

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

ALGORITHM 3.1. For a given $J^*(\boldsymbol{\theta})$ that satisfies Assumptions ID and Jac, 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 nonzero $d_{\boldsymbol{\mu}} \times 1$ vector $m^{(1)}$ such that when $\beta = 0$,

$$J^*(\boldsymbol{\theta}) m^{(1)}(\boldsymbol{\mu}) = 0_{d_g \times 1} \tag{3.3}$$

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

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 $h^{(1)} : \mathcal{M}^{(1)} \to \boldsymbol{\mathcal{M}}$ 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 $\mu^{(1)} \in \mathcal{M}^{(1)}$.[11]

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

[11] 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 $\partial h^{(1)}(\mu^{(1)})/\partial \mu^{(1)\prime}$.

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

$$\frac{\partial g^{(i-1)}(\theta^{(i-1)})}{\partial \mu^{(i-1)'}} m^{(i)}(\mu^{(i-1)}) = 0_{d_g \times 1} \tag{3.5}$$

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

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:

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

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

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 $\partial h^{(i)}(\mu^{(i)})/\partial \mu^{(i)'}$ has full rank and (2)

$$\frac{\partial h^{(i)}(\mu^{(i)})}{\partial(\pi_1^{(i)}, \ldots, \pi_{i-1}^{(i)})} = \begin{bmatrix} 0_{(d_{\boldsymbol{\mu}} - d_\pi) \times (i-1)} \\ C^{(i)}(\mu^{(i)}) \\ 0_{(d_\pi - i + 1) \times (i-1)} \end{bmatrix},$$

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

8. If $i = d_\pi$, stop and set $h = h^{(1)} \circ \cdots \circ h^{(d_\pi)}$ and $\mu = \mu^{(d_\pi)}$. Otherwise, set $\theta^{(i)} = (\beta, \mu^{(i)})$, $g^{(i)}(\theta^{(i)}) = g^{(i-1)}(\beta, h^{(i)}(\mu^{(i)}))$, $\Theta^{(i)} = \{(\beta, \mu^{(i)}) \in \mathbb{R}^{d_{\boldsymbol{\theta}}} : (\beta, h^{(i)}(\mu^{(i)})) \in \Theta^{(i-1)}\}$ 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), that is, $\boldsymbol{\mu} = h(\mu)$. 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 3.1. Defining the matrix function $M^{(i)}(h^{(i)}(\mu^{(i)})) = \partial h^{(i)}(\mu^{(i)})/\partial \mu^{(i)'}$ 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

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

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 unique though such nonuniqueness 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 $i = 1, \ldots, d_\pi$ with $\mathcal{M}^{(0)} \equiv \boldsymbol{\mathcal{M}}$.*

PROPOSITION 3.1. *Under Assumptions ID, Jac, 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 $\boldsymbol{g}^*(\boldsymbol{\theta})$. In practice, one can verify this assumption by simply calculating $m^{(i)}(\cdot)$ in Steps 2 or 5 in Algorithm 3.1, as these steps are straightforward to implement.

REMARK 3.2. In some cases, it may be difficult to solve the ODEs (3.4) analytically. Nevertheless, an abundance of numerical methods for solving systems of ODEs is readily available in the literature.[12] One can numerically solve for the $h^{(i)}$ functions in (3.4) using methods in, for example, Quarteroni, Sacco, and Saleri (2010). To summarize a standard approach to this problem, one can first approximate the $h^{(i)}$ functions using known basis functions and transform the system of ODEs into a nonlinear system of equations. Then a Newton–Raphson-type method can be implemented to solve for the coefficients on the basis functions, thus obtaining a numerical solution for the $h^{(i)}$ functions. See Chapters 7–11 of Quarteroni, Sacco, and Saleri (2010) for details on the choice of basis functions and algorithms used to implement this approach as well as other details for numerical methods with nonlinear ODEs.

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) and Choi and Phillips (1992) 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 underidentified across its entire parameter space (not just in the subspace 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 models that are globally underidentified across their

[12] We thank Andres Santos for pointing this out.

parameter space 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 underidentified 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_ζ elements of $h(\mu)$ are equal to the first d_ζ elements of the new well-identified parameter $\zeta = (\zeta^1, \zeta^2)$, namely, $\boldsymbol{\zeta} = (h_1(\mu), \ldots, h_{d_\zeta}(\mu)) = \zeta^1$. This is the special case described in Remark 2.1. 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 ζ is the new always well identified parameter.

We summarize the notation for the original and transformed parameters in Table 1 and close this section by illustrating the reparameterization algorithm with two other examples discussed earlier.

EXAMPLE 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 $\boldsymbol{g}^*(\boldsymbol{\theta}) = \xi^* - \mathfrak{g}(\boldsymbol{\theta})$, where ξ^* is equal to a vector of the probabilities $p_{yd,z}$'s and $\mathfrak{g}(\boldsymbol{\theta})$ is defined in (2.5).[13] The Jacobian for this model with respect to $\boldsymbol{\mu}$ is

$$\begin{aligned} J^*(\boldsymbol{\theta}) &= -\frac{\partial \mathfrak{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}'} \\ &= -\begin{bmatrix} C_2(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & 0 & C_1(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & C_3(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \\ C_2(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & 0 & C_1(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & C_3(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \\ -C_2(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & 1 - C_1(\boldsymbol{\pi}_1, \zeta; \boldsymbol{\pi}_3) & 0 & -C_3(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \\ -C_2(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & 1 - C_1(\boldsymbol{\pi}_1, \zeta; \boldsymbol{\pi}_3) & 0 & -C_3(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \\ 1 - C_2(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & 0 & -C_1(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & -C_3(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \\ 1 - C_2(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & 0 & -C_1(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) & -C_3(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3) \end{bmatrix} \end{aligned}$$

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 $m(\boldsymbol{\mu}) = (0, C_3(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3)/(1 -$

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

$C_1(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3)), -C_3(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3)/C_1(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3), 1)'$. For Step 2, a set of general solutions to the system of ODEs,

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

is implied by

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

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

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

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 $h(\cdot)$:

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

where $a(\mu) = \pi(1-\zeta_1)$, $b(\mu) = (1-\zeta_1)(1-\pi-\pi\zeta_2)$ and $c(\mu) = \zeta_2[\pi(1-\zeta_1)-1]$.[14] For any choice of copula, we can also express the new parameters as a function of the original ones as follows:

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

EXAMPLE 2.4 (Continued). In this example, we again consider GMM estimation so that $\boldsymbol{g}^*(\boldsymbol{\theta}) = E_{\gamma^*}\varphi(W_i, \boldsymbol{\theta})$, where the moment function $\varphi(w, \boldsymbol{\theta})$ is given by (2.7). The Jacobian with respect to $\boldsymbol{\mu}$ is

$$J^*(\boldsymbol{\theta}) = -E_{\gamma^*}A(Y_{h,i})\begin{bmatrix} \boldsymbol{\pi}_2 + \boldsymbol{\pi}_3 & \boldsymbol{\pi}_1 & \boldsymbol{\pi}_1 \\ -\boldsymbol{\pi}_2 & 1-\boldsymbol{\pi}_1 & 0 \end{bmatrix}$$

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:

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

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

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

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, that is, 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

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

being full rank. Thus, we have found a reparameterization function $h(\cdot)$ satisfying the conditions 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, $\mu = (\zeta_1, \zeta_2, \pi) = (\log(\boldsymbol{\pi}_2(1-\boldsymbol{\pi}_1)), \boldsymbol{\pi}_1(\boldsymbol{\pi}_2 + \boldsymbol{\pi}_3), \log((1-\boldsymbol{\pi}_1)/\boldsymbol{\pi}_1))$.

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

The nonuniqueness of the reparameterizations resulting from Procedure 3.1 or Algorithm 3.1 leads to the natural question of how to "choose" the appropriate reparameterization in practice. A natural criterion for guidance on the choice of reparameterization is interpretability. With this criterion in mind, we recommend choosing a reparameterization function $h(\cdot)$ such that, after suitable permutation of its rows and columns, $\partial h(\mu)/\partial \mu'$ is triangular. The inverse function theorem implies that (after suitable permutations) $\partial h^{-1}(\boldsymbol{\mu})/\partial \boldsymbol{\mu}'$ will similarly be a triangular matrix. This would then imply, for example, that ζ_1 is a function of only one element of $\boldsymbol{\mu}$, ζ_2 is a function of two or less elements of $\boldsymbol{\mu}$, and so on. This is desirable from an interpretability perspective as it allows the user to understand which functions of the original model parameters are identified, with the goal of keeping these functions as simple as possible. An additional way of doing this is to choose the reparameterization function so that $\partial h(\mu)/\partial \mu'$ is sparse. Interpretability is likely to be easiest when one chooses $h(\cdot)$ to have both of these properties. In practice, the "choice" of reparameterization is essentially dictated by the choice of constants of integration when moving from the general solution to a system of ODEs to a particular solution, in, for example, steps 3 and 7 of Algorithm 3.1. Note that in the illustrative reparameterizations given for Examples 2.1, 2.3, and 2.4 above, we indeed choose the constants of integration ($c_i(\zeta)$'s and $c^i(\zeta)$'s) following the general advice given here.

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 reparametrized 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, semistrong, and weak (which includes nonidentification). 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 subset of $\mathbb{R}^{d_{\boldsymbol{\theta}}}$[15] 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 $i, m \geq 1$.[16] We pause here to illustrate the form of this parameter space in two of our running examples.

[15]Technically, the *true* parameter space for $\boldsymbol{\theta}$ must be a subspace of the interior of the *optimization* parameter space $\boldsymbol{\Theta}$. We suppress this distinction in the main text for ease of notation and refer the interested reader to the Appendix for further details.

[16]Technically, there must exist a metric on Φ^* such that (W_i, W_{i+m}) under γ converges in distribution to (W_i, W_{i+m}) under γ_0 for any $\gamma \to \gamma_0$.

EXAMPLES 2.1 and 2.2 (Continued). We again focus upon Example 2.1 since the parameter space for Example 2.2 has similar features. Here, we also focus on the classic case for which $F_{\varepsilon\nu}$ is a bivariate standard normal distribution. The parameter space for $\boldsymbol{\theta}$ takes the form $\boldsymbol{\Theta} = \mathcal{B} \times \boldsymbol{\mathcal{Z}} \times \boldsymbol{\Pi}^1 \times \boldsymbol{\Pi}^2$, where $\mathcal{B} = \bigtimes_{j=1}^{l-1}[b_{L,j}, b_{H,j}]$ with $b_{L,j}, b_{H,j} \in \mathbb{R}$ such that $-1 < b_{L,j} \leq 0 \leq b_{H,j} < 1$ and $b_{L,j} \neq b_{H,j}$ for $j = 1, \ldots, d_\beta$, $\boldsymbol{\mathcal{Z}} \subset \mathbb{R}$, $\boldsymbol{\Pi}^1 \subset \mathbb{R}^k$, $\boldsymbol{\Pi}^2 \subset (-1, 1)$, $\boldsymbol{\mathcal{Z}}$, $\boldsymbol{\Pi}^1$ and $\boldsymbol{\Pi}^2$ are compact. In this example, ϕ denotes the joint distribution of $(X_i, Z_i, \varepsilon_i, \nu_i)$. Define $a_i(\beta, \boldsymbol{\zeta}) \equiv (X_i', q_i)'$ and let P_ϕ and E_ϕ denote probability and expectation under ϕ. Then we can define the parameter space for ϕ as follows:

$$
\begin{aligned}
\boldsymbol{\Phi}(\boldsymbol{\theta}) = \Big\{ & \phi \in \Phi^* : E_\phi[\varepsilon_i \nu_i] = \boldsymbol{\pi}^2, \\
& P_\phi\big(Z_i' c = 0\big) < 1 \text{ for any } c \neq 0, E_\phi\big[\|Z_i\|^{4+\varepsilon} + \|X_i\|^{4+\varepsilon}\big] \leq C, \\
& E_\phi \sup_{(\beta,\boldsymbol{\zeta}) \in \mathcal{B} \times \boldsymbol{\mathcal{Z}}} \big[q\big(\boldsymbol{\zeta} + Z_{1i}'\beta\big)^{4+\varepsilon} + F_\nu^{-1}\big(-\boldsymbol{\zeta} - Z_{1i}'\beta\big)^{4+\varepsilon} + \big|L_i(\beta, \boldsymbol{\zeta})\big|^{2+\varepsilon}\big] \leq C, \\
& E_\phi\big[D_i a_i(\beta, \boldsymbol{\zeta}) a_i(\beta, \boldsymbol{\zeta})'\big] \in \mathbb{R}^{(k+1)\times(k+1)} \text{ has full rank for all} \\
& (\beta, \boldsymbol{\zeta}) \in \mathcal{B} \times \boldsymbol{\mathcal{Z}} \text{ with } \beta \neq 0 \Big\},
\end{aligned}
$$

for some constants $C < \infty$ and $\varepsilon > 0$.

EXAMPLE 2.3 (Continued). The parameter space for $\boldsymbol{\theta}$ takes the form

$$
\boldsymbol{\Theta} = \big\{ \boldsymbol{\theta} = (\beta, \boldsymbol{\zeta}, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3) \in [b_L, b_H] \times \boldsymbol{\mathcal{Z}} \times \boldsymbol{\Pi} : t_L \leq \beta + \boldsymbol{\zeta} \leq t_H \big\},
$$

where $b_L, b_H, t_L, t_H \in \mathbb{R}$ such that $-1 < b_L \leq 0 \leq b_H < 1$ with $b_L \neq b_H$, $0 < t_L < t_H < 1$, $\boldsymbol{\mathcal{Z}} \subseteq [t_l, t_h]$, $\boldsymbol{\Pi} \subset \mathbb{R}^3$, $\boldsymbol{\mathcal{Z}}$ and $\boldsymbol{\Pi}$ are compact. In this example, $\phi = P_\gamma(Z_i = 1)$ and $\boldsymbol{\Phi}(\boldsymbol{\theta})$ does not depend upon $\boldsymbol{\theta}$ so that this parameter space can be defined as $\boldsymbol{\Phi} = [p_L, p_H]$, where $0 < p_L < p_H < 1$.

The next lemma formally establishes the properties of the reparameterization function $h(\cdot)$.

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.[17]

LEMMA 4.1. *Under Assumptions* ID, Jac, *and* H, (i) *the function* $h : \mathcal{M} \to \boldsymbol{\mathcal{M}}$ *is a homeomorphism, and hence bijective*; (ii) $h(\mu)$ *is continuously differentiable on* $\mathcal{M}$.

[17]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.

Due to this result, we can equivalently derive limit theory under sequences of parameters in $\mathbf{\Gamma}$ or in the following transformed parameter space:

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

where $\Phi^*(\theta) \equiv \mathbf{\Phi}^*(\beta, h(\mu)) \subset \Phi^*$ for all $\theta \in \boldsymbol{\Theta}$.[18] Define sets of sequences of parameters $\{\gamma_n\}$ as follows:

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

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

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

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 nonidentification 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 $(\hat{\beta}_n, h(\hat{\mu}_n))$ 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 $\operatorname{rank}(h_\pi^s(\mu)) = \tilde{d}_\pi^*$ for all $\mu \in \mathcal{M}_\epsilon \equiv \{\mu : (\beta, \mu) \in \boldsymbol{\Theta}, \|\beta\| < \epsilon\}$ 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), Choi and Phillips (1992), Sims, Stock, and Watson (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. $\operatorname{rank}(h_\pi^s(\mu)) = \tilde{d}_\pi^*$ *for some constant* $\tilde{d}_\pi^* \leq d_\pi$ *for all* $\mu \in \mathcal{M}_\epsilon$ *for some* $\epsilon > 0$.

[18] In analogy to the remark made in footnote 15, the *true* parameter space for the transformed parameter θ actually differs from $\boldsymbol{\Theta}$ but the difference is suppressed in the main text for ease of notation. We again refer the interested reader to the Appendix for details.

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, for example, Assumptions R1 and R2 of AC14. With an explicit $h(\cdot)$ found, for example, 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 nonlinear and one wishes to obtain the joint limiting behavior of a larger subvector of $\hat{\boldsymbol{\mu}}_n$ 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 entries 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).

Example 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 $\hat{\boldsymbol{\mu}}^s_n = (\hat{\boldsymbol{\mu}}_{n,3}, \hat{\boldsymbol{\mu}}_{n,4})$. In this case under $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$,

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

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

$$\tilde{\eta}_n(\hat{\mu}_n) = \sqrt{n}\mathcal{S}(\hat{\mu}_n)\begin{bmatrix} \tilde{\eta}_n^N(\hat{\mu}_n) \\ \tilde{\eta}_n^D(\hat{\mu}_n) \end{bmatrix}(\hat{\pi}_n - \pi_n),$$

where

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

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

and $\mathcal{S}(\hat{\mu}_n) = O_p(1)$, which we obtain by using the results from Lemma A.1 in the Online Appendix B. (The derivations behind the above expressions can be found in the Online Appendix B as well.) 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* ID, CF, Reg1, Jac, H, Reg2 *and* Reg3, *and Assumptions* B1–B3 *and* C1–C6 *of AC*12,[19] *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

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

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

with $\pi_{0,b}^*$ *being a random vector that minimizes a noncentral 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* $d_{\boldsymbol{\mu}} - d_\pi$ *entries. The underlying Gaussian process* $G_0(\cdot) \equiv G(\cdot; \gamma_0)$ *is defined in Assumption C*3 *of AC*12 *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 AC*12, *respectively.*

[19]Here and below, we refer the reader to the Appendix for the assumptions of AC12. In the Online Appendix B, we also provide sufficient conditions for all the assumptions used in this paper including those from AC12 for the threshold crossing model (Example 2.3).

(ii) *Suppose Assumptions* ID, CF, Reg1, Jac, H, Reg2 *and* Reg3, *and Assumptions* B1–B3, C1–C5, C7–C8, *and* D1–D3 *of AC*12, *applied to the* θ *and* $Q_n(\theta)$ *of this paper, hold. Under parameter sequences* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

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

if $\beta_0 = 0$ *and*

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

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 AC*12.

Theorem 4.1 describes the joint limiting behavior of $\hat{\beta}_n$ and $\hat{\boldsymbol{\mu}}_n^s$ under a comprehensive class of identification strengths. By rotating the subvector $\hat{\boldsymbol{\mu}}_n^s$ 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 $\hat{\boldsymbol{\mu}}_n$.

REMARK 4.1. Note that the results of Theorem 4.1 hold regardless of the choice of reparameterization function $h(\cdot)$, as long as it satisfies the imposed assumptions. This implies that the limiting random variables given in the theorem are *invariant to the choice of reparameterization*. Although this statement may seem contradictory because the reparameterization function $h(\cdot)$ appears in the expressions describing the random variables, note also that the objects ζ_0 and π that also appear in these expressions are *different for different choices of reparameterization*. An analogous version of this remark similarly applies to Corollary 4.1 below.

Though nonlinearity of the reparameterization function often makes it impossible to obtain a $\sqrt{n}$-consistent rotation of the full vector $\hat{\boldsymbol{\mu}}_n$ 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 $\boldsymbol{\mu} = h(\zeta, \pi)$ according to whether they depend upon π or not. Without loss of generality, suppose that the first $d_{\boldsymbol{\mu}^1}$ components of $h(\zeta, \pi)$ do not actually depend upon π (e.g., in cases described by Remark 3.5), while the final $d_{\boldsymbol{\mu}} - d_{\boldsymbol{\mu}^1}$ of $h(\zeta, \pi)$ do. Denote the corresponding entries of $\boldsymbol{\mu} = h(\zeta, \pi)$ as $\boldsymbol{\mu}^1 = h^1(\zeta)$ and $\boldsymbol{\mu}^2 = h^2(\zeta, \pi)$, 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 reparametrized parameters θ, namely,

$$H_0 : r(\theta) \equiv \boldsymbol{r}\big(\beta, h(\mu)\big) = v \in \mathbb{R}^{d_{\boldsymbol{r}}},$$

With this notation in mind, a standard Wald statistic for H_0 based upon $\hat{\boldsymbol{\theta}}_n = (\hat{\beta}_n, h(\hat{\mu}_n))$ can be written as[20]

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

[20]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 $(1, h_{\mu}(\hat{\mu}_n))B^{-1}(\hat{\beta}_n)\hat{\Sigma}_n B^{-1}(\hat{\beta}_n)(1, h_{\mu}(\hat{\mu}_n))'$.

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 of AC14 and V1–V2 of AC12, 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 underidentification 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) underidentification. 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) $\operatorname{rank}(r_\pi(\theta)) = d_\pi^*$ *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 $(\beta, h(\mu))$ is continuously differentiable on Θ by Lemma 4.1(ii). Since $(1, h_\mu(\mu))$ 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 $\partial \boldsymbol{r}(\beta, h(\mu))/\partial \boldsymbol{\mu}'$ 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

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

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, that is, $H_0 : \boldsymbol{\theta}^s = v$ for some subvector $\boldsymbol{\theta}^s$ of $\boldsymbol{\theta}$, Assumption R2 is equivalent to Assumption Reg3 introduced in the

previous section.[21] 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).[22]

ASSUMPTION $\mathrm{R_L}$. *$r(\theta) = R\theta$, where R is a $d_r \times d_\theta$ matrix with full row rank.*

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

[21] 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.

[22] These statements hold because β is not a function of π.

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 AC*12, *applied to the* θ *and* $Q_n(\theta)$ *of this paper, hold. Under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $\|b\| < \infty$,

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

where $\{\lambda(\pi; \gamma_0, b) : \pi \in \Pi\}$ *is a stochastic process defined in the Appendix.*

(ii) *Suppose Assumptions* CF, ID, Reg1, Jac, H, R1 *and* R2, *and Assumptions* B1–B3, C1–C5, C7–C8, D1–D3 *and* V2 *of AC*12, *applied to the* θ *and* $Q_n(\theta)$ *of this paper, hold. Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$,

$$W_n(v_n) \xrightarrow{d} \chi^2_{d_r}.$$

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 (Forthcoming) 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 (Forthcoming) and Andrews and Mikusheva (2016a).[23] There is an interesting complementarity here between our results and those of Andrews and Guggenberger (Forthcoming) 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

[23] 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.

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 (Forthcoming) and Andrews and Mikusheva (2016a) do. We should also note that the frameworks of our paper and Andrews and Guggenberger (Forthcoming) or Andrews and Mikusheva (2016a) are nonnested and a key limiting feature of our approach that is not present in any of the other papers mentioned in this remark is that our approach applies only to models for which a parameter governs identification strength.

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 ours 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.[24]

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* $\delta_0 \in \mathbb{R}^{d_\delta}_\infty$ *such that under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *or* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$ *there is an estimator* $\hat{\delta}_n$ *with* $\hat{\delta}_n \xrightarrow{p} \delta_0$.

Given the above assumption, we may now redefine $c_{1-\alpha}(\cdot)$ to be a function of only the finite-dimensional parameter $(b, \zeta_0, \pi_0, \delta_0)$. That is, let $\ell_0 = (b, \zeta_0, \pi_0, \delta_0)$ and let $c_{1-\alpha}(\ell_0)$ denote the $1-\alpha$ quantile of $\lambda(\pi^*_{0,b}; \gamma_0, b)$. 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 because it accounts for the fact that in any finite sample, using the

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

data to determine identification strength comes with a positive probability of miscategorizing the identification strength. Neglecting this fact can induce overrejection of the null hypothesis in finite samples.

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:

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

where $\chi_{d_r}^2 (1-\alpha)^{-1}$ is the $(1-\alpha)$ quantile of a $\chi_{d_r}^2$-distributed random variable and $c_{1-\alpha,n}^{\mathrm{LF}} \equiv \sup_{\ell \in \hat{\mathcal{L}}_n \cap \mathcal{L}(v)} c_{1-\alpha}(\ell)$ with $\hat{\mathcal{L}}_n \equiv \{\ell = (b, \zeta, \pi, \delta) \in \mathcal{L} : (\zeta, \delta) = (\hat{\zeta}_n, \hat{\delta}_n)\}$, $\mathcal{L}(v) \equiv \{\ell \in \mathcal{L} : r(\theta) = v\}$, and $\mathcal{L} \equiv \{\ell = (b, \zeta, \pi, \delta) \in \mathbb{R}_\infty^{d_\beta + d_\zeta + d_\pi + d_\delta}$: there is some $\gamma_0 \in \Gamma$ such that $\zeta = \zeta_0$, $\pi = \pi_0$, $\delta = \delta_0$ and 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).[25] 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* $\mathcal{L}_0 \equiv \{\ell = (b, \zeta, \pi, \delta) \in \mathcal{L} : (\zeta, \delta) = (\zeta_0, \delta_0)\}$.

This assumption is assured to hold, for example, 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 AC*12 *and Assumptions* FD *and* DF1, $\limsup_{n\to\infty} \sup_{\gamma \in \Gamma : r(\theta) = v} P_\gamma(W_n(v) > c_{1-\alpha,n}^{\mathrm{ICS}}) = \alpha$.

This proposition shows that for certain, namely low-dimensional, null hypotheses, a test that rejects when the standard Wald statistic $W_n(v)$ exceeds the ICS CV $c_{1-\alpha,n}^{\mathrm{ICS}}$ described in this section has correct asymptotic size.

[25] 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 $c_{1-\alpha}^{\mathrm{LF}}$. 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 $c_{1-\alpha}(\ell)$ 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 $c_{1-\alpha}(\ell)$ 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.[26] 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 $\hat{I}_n^a(\hat{b}_n, \hat{\pi}_n) = \{(b,\pi) \in \mathcal{P}(\hat{\zeta}_n, \hat{\delta}_n) : [(\hat{b}_n - b)', (\hat{\pi}_n - \pi)']\hat{\tilde{\Sigma}}_n^{-1}[(\hat{b}_n - b)', (\hat{\pi}_n - \pi)']' \le \chi^2_{d_\beta+d_\pi}(1-a)^{-1}\}$, where

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

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: $\hat{I}_n^a(\hat{b}_n, \hat{\pi}_n) = \{(b,\pi) \in \mathcal{P}(\hat{\zeta}_n, \hat{\delta}_n) : (\hat{b}_n - b)'\hat{\Sigma}_{\beta\beta,n}^{-1}(\hat{b}_n - b) \le \chi^2_{d_\beta}(1-a)^{-1}\}$. 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 $\|b\| \to \infty$. Similar 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 $\tilde{\mathcal{L}}_n^a(b, \gamma_0) = \{\ell = (\tilde{b}, \zeta, \pi, \delta) \in \hat{\mathcal{L}}_n : (\tilde{b}, \pi) \in \hat{I}_n^a(b + \tau_{0,b}^{\beta}(\pi_{0,b}^*), \pi_{0,b}^*)\}$ and $\hat{\mathcal{L}}_n^a = \{\ell = (b, \zeta, \pi, \delta) \in \hat{\mathcal{L}}_n : (b,\pi) \in \hat{I}_n^a(\hat{b}_n, \hat{\pi}_n)\}$. For a size-$\alpha$ test, the construction of the CV proceeds in two steps:

[26] A similar complication arises in, for example, 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 $\mathfrak{s} = \mathfrak{s}(\hat{\zeta}_n, \hat{\delta}_n, \hat{\bar{\Sigma}}_n)$ such that

$$\sup_{(b,\gamma_0)\in\mathbb{R}_\infty^{d_\beta}\times\Gamma:(b,\zeta_0,\pi_0,\delta_0)\in\hat{\mathcal{L}}_n\cap\mathcal{L}(v)} P\Big(\lambda\big(\pi^*_{0,b};\gamma_0,b\big)\geq \sup_{\ell\in\tilde{\mathcal{L}}_n^a(b,\gamma_0)\cap\mathcal{L}(v)} c_{1-\alpha}(\ell)+\mathfrak{s}\Big)\leq\alpha.$$

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

By Assumption FD, the computations in Step 1 can be achieved by simulating from the joint distribution of $\lambda(\pi^*_{0,b};\gamma_0,b)$, $\tau^\beta_{0,b}(\pi^*_{0,b})$, and $\pi^*_{0,b}$ over a grid of (b,γ_0) values such that $(b,\zeta_0,\pi_0,\delta_0)\in\hat{\mathcal{L}}_n\cap\mathcal{L}(v)$ or by using more computationally efficient global optimization methods such as response surface analysis (see, e.g., Jones, Schonlau, and Welch (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 $\mathcal{L}_0^a(b,\gamma_0) = \{\ell = (\tilde{b},\zeta,\pi,\delta)\in\mathcal{L}_0 : (\tilde{b},\pi)\in I_0^a(b+\tau^\beta_{0,b}(\pi^*_{0,b}),\pi^*_{0,b})\}$. When using the first $\hat{I}_n^a(\hat{b}_n,\hat{\pi}_n)$ described above,

$$\begin{aligned}
&I_0^a\big(b+\tau^\beta_{0,b}(\pi^*_{0,b}),\pi^*_{0,b}\big)\\
&\quad=\big\{(b,\pi)\in\mathcal{P}(\zeta_0,\delta_0):\\
&\qquad\big[(\tau^\beta_{0,b}(\pi^*_{0,b})',(\pi^*_{0,b}-\pi)'\big]\bar{\Sigma}_0^{-1}\big(b+\tau^\beta_{0,b}(\pi^*_{0,b}),\theta^*_{0,b}\big)\big[(\tau^\beta_{0,b}(\pi^*_{0,b})',(\pi^*_{0,b}-\pi)'\big]'\\
&\qquad\leq\chi^2_{d_\beta+d_\pi}(1-a)^{-1}\big\}
\end{aligned}$$

with

$$\begin{aligned}
&\bar{\Sigma}_0\big(b+\tau^\beta_{0,b}(\pi^*_{0,b}),\theta^*_{0,b}\big)\\
&\quad\equiv\begin{pmatrix}\Sigma_{\beta\beta,0}(\theta^*_{0,b}) & \big\|b+\tau^\beta_{0,b}(\pi^*_{0,b})\big\|^{-1}\Sigma_{\beta\pi,0}(\theta^*_{0,b})\\ \big\|b+\tau^\beta_{0,b}(\pi^*_{0,b})\big\|^{-1}\Sigma_{\beta\pi,0}(\theta^*_{0,b})' & \big\|b+\tau^\beta_{0,b}(\pi^*_{0,b})\big\|^{-2}\Sigma_{\pi\pi,0}(\theta^*_{0,b})\end{pmatrix}
\end{aligned}$$

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 Assumption V1 of AC12.) When using the second $\hat{I}_n^a(\hat{b}_n,\hat{\pi}_n)$ described above,

$$\begin{aligned}
&I_0^a\big(b+\tau^\beta_{0,b}(\pi^*_{0,b}),\pi^*_{0,b}\big)\\
&\quad=\big\{(b,\pi)\in\mathcal{P}(\zeta_0,\delta_0):\tau^\beta_{0,b}(\pi^*_{0,b})'\Sigma^{-1}_{\beta\beta,0}(\theta^*_{0,b})\tau^\beta_{0,b}(\pi^*_{0,b})\leq\chi^2_{d_\beta}(1-a)^{-1}\big\}.
\end{aligned}$$

Assumption DF2. *There exists some* $(b^*,\gamma_0^*)\in\mathbb{R}_\infty^{d_\beta}\times\Gamma$ *such that for some* $\{\gamma_n\}\in\Gamma(\gamma_0^*)$, $n^{1/2}\beta_n\to b^*$ *and*:

(i) $P(\lambda(\pi^*_{0,b^*}; \gamma^*_0, b^*) \geq \sup_{\ell \in \mathcal{L}^a_0(b^*, \gamma^*_0) \cap \mathcal{L}(v)} c_{1-\alpha}(\ell) + s(\zeta^*_0, \delta^*_0, \bar{\Sigma}(b^*, \gamma^*_0))) = \alpha$,

(ii) $P(\lambda(\pi^*_{0,b^*}; \gamma^*_0, b^*) = \sup_{\ell \in \mathcal{L}^a_0(b^*, \gamma^*_0) \cap \mathcal{L}(v)} c_{1-\alpha}(\ell) + s(\zeta^*_0, \delta^*_0, \bar{\Sigma}(b^*, \gamma^*_0))) = 0$.

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,* $\limsup_{n\to\infty} \sup_{\gamma \in \Gamma : r(\theta) = v} P_\gamma(W_n(v) > c^{\mathrm{AB}}_{1-\alpha,n}) = \alpha$.

Similar to Proposition 6.1, this proposition shows that for certain null hypotheses of interest, a test that rejects when the Wald statistic $W_n(v)$ exceeds the adjusted-Bonferroni CV $c^{\mathrm{AB}}_{1-\alpha,n}$, whose construction is detailed in this section, has correct asymptotic size.

REMARK 6.1. In combination with the standard Wald statistic, the two CV construction methods detailed in this section allow one to conduct subvector inference for subvectors of the model parameters $\boldsymbol{\theta}$. One of the main alternative approaches to subvector inference with proven asymptotic size control would be to use an existing identification-robust test *for the entire parameter vector* $\boldsymbol{\theta}$, such as those advanced by Andrews and Guggenberger (Forthcoming) and Andrews and Mikusheva (2016a), and "project the test" onto the space generated by the lower-dimensional null hypothesis. Apart from their well-known poor power properties when the dimension of the full parameter vector is substantially larger than the subvector under test, projection-based tests can also be computationally costly. The computation of a projection-based typically involves minimizing a nonlinear test statistic function over the space of nuisance parameters that are not under a test. Our CV constructions also require the optimization of a nonlinear function $c_{1-\alpha}(\ell)$ but the dimension of the space over which the optimization is performed is equal to $d_\beta + d_\pi$, which is typically of lower dimension than the nuisance parameters, at least for low-dimensional hypotheses. Nevertheless, $c_{1-\alpha}(\ell)$ must typically be computed via Monte Carlo simulation. So there is no general rule for which of the two tests is easier to compute.

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 have 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/or endogenous regressor are/is binary (e.g., Evans and Schwab (1995), Goldman, Bhattacharya, Mccaffrey, Duan, Leibowitz, Joyce, and Morton (2001), Lochner and Moretti (2004), Altonji, Elder, and Taber (2005), Rhine, Greene, and Toussaint-Comeau (2006)) and instruments are potentially weak. This section contains some of the objects that appear in the general results of this paper, applied to the threshold crossing

model. The full verification of the assumptions imposed in the theoretical results of this paper and other technical details are contained in Online Appendix B.

The random sample is given by the vector $W_i \equiv (Y_i, D_i, Z_i)$ for $i = 1, \ldots, n$. 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 $\boldsymbol{\Theta} \equiv \{\boldsymbol{\theta} = (\beta, \boldsymbol{\zeta}, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \pi) \in [-0.98 - \boldsymbol{\epsilon}, 0.98 + \boldsymbol{\epsilon}] \times [0.01 - \boldsymbol{\epsilon}, 0.99 + \boldsymbol{\epsilon}] \times [0.01 - \boldsymbol{\epsilon}, 0.99 + \boldsymbol{\epsilon}] \times [0.01 - \boldsymbol{\epsilon}, 0.99 + \boldsymbol{\epsilon}] \times [-0.99 - \boldsymbol{\epsilon}, 0.99 + \boldsymbol{\epsilon}] : 0.01 - \boldsymbol{\epsilon} \leq \beta + \boldsymbol{\zeta} \leq 0.99 + \boldsymbol{\epsilon}\}$:

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

for $\boldsymbol{\epsilon} = 0.005$, where $\boldsymbol{\rho}(w, \boldsymbol{\theta})$ is a function of $\mathfrak{g}(\boldsymbol{\theta})$ via $\boldsymbol{\rho}(w, \boldsymbol{\theta}) \equiv -\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(w) \times \log \boldsymbol{p}_{yd,z}(\boldsymbol{\theta})$, the logarithm of density function[27] 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). The parameter space $\boldsymbol{\Theta}$ is chosen here to be unrestrictive while still satisfying the conditions imposed in Online Appendix B. These latter conditions are mainly imposed to avoid boundary issues in the estimation of $\boldsymbol{\theta}$. See Online Appendix B and AC12 for further details.

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 Online Appendix B.

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

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

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

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

[27] 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.

where $\phi_{z,0} \equiv P_{\gamma_0}(Z_i = z)$, $\psi \equiv (\beta, \zeta)$, $\psi_0 \equiv (0, \zeta_0)$ 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

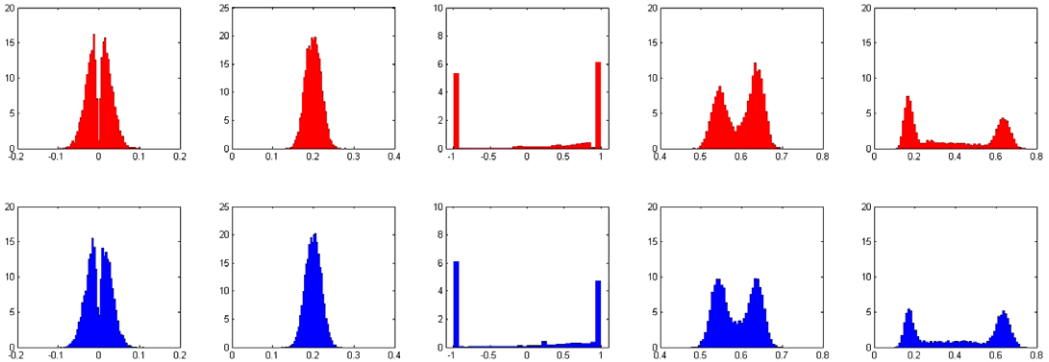
$$\Omega(\pi_1, \pi_2; \gamma_0) = S_\psi V^\dagger\big((\psi_0, \pi_1), (\psi_0, \pi_2); \gamma_0\big) S_\psi',$$

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

$$\begin{aligned}
&V^\dagger(\theta_1, \theta_2; \gamma_0) \\
&\quad \equiv E_{\gamma_0}\left[\left(\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i) \frac{D_\theta p^\dagger_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)}\right)\left(\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i) \frac{D_\theta p^\dagger_{yd,z}(\theta_2)'}{p_{yd,z}(\theta_2)}\right)\right] \\
&\qquad - E_{\gamma_0}\left[\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i) \frac{D_\theta p^\dagger_{yd,z}(\theta_1)}{p_{yd,z}(\theta_1)}\right] E_{\gamma_0}\left[\sum_{y,d,z=0,1} \mathbf{1}_{ydz}(W_i) \frac{D_\theta p^\dagger_{yd,z}(\theta_2)'}{p_{yd,z}(\theta_2)}\right] \\
&\quad = \sum_{y,d,z=0,1} \frac{p_{yd,z}(\theta_0)\phi_{z,0}}{p_{yd,z}(\theta_1) p_{yd,z}(\theta_2)} D_\theta p^\dagger_{yd,z}(\theta_1) p^\dagger_{yd,z}(\theta_2)' \\
&\qquad - \left(\sum_{y,d,z=0,1} \frac{p_{yd,z}(\theta_0)\phi_{z,0}}{p_{yd,z}(\theta_1)} D_\theta p^\dagger_{yd,z}(\theta_1)\right)\left(\sum_{y,d,z=0,1} \frac{p_{yd,z}(\theta_0)\phi_{z,0}}{p_{yd,z}(\theta_2)} D_\theta p^\dagger_{yd,z}(\theta_2)'\right)
\end{aligned}$$

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

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 1–3 provide the simulated finite-sample density functions of the estimators of the threshold-crossing model parameters in the top row and their asymptotic approximations in the bottom row. 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 $\beta_0 + \tau^\beta_{0,b}(\pi^*_{0,b})/\sqrt{n}$ and $h(\zeta_0 + \tau^\zeta_{0,b}(\pi^*_{0,b})/\sqrt{n}, \pi^*_{0,b})$ 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 $\sqrt{n}$-consistent parameter estimate $\hat{\zeta}_n$ in the distribution of $\hat{\boldsymbol{\mu}}_n$. Figures 1–3 show that (i) the distributions of the parameter estimators can be highly non-Gaussian under weak/nonidentification; (ii) as β grows larger, the distributions become approximately



Asymptotic (bottom row) and finite-sample (top row, $n = 1000$) 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 $\pi = (0.6, 0.4, 0.4)$.

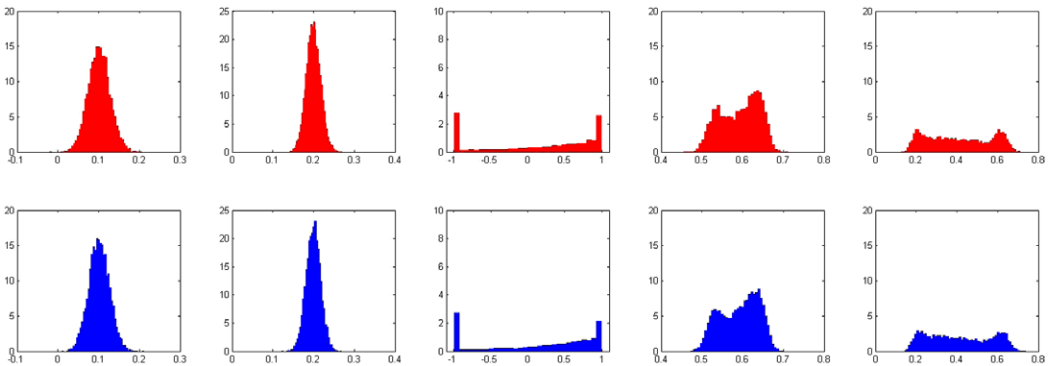
FIGURE 1. Threshold crossing model parameter estimator densities when $b = 0$.

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*

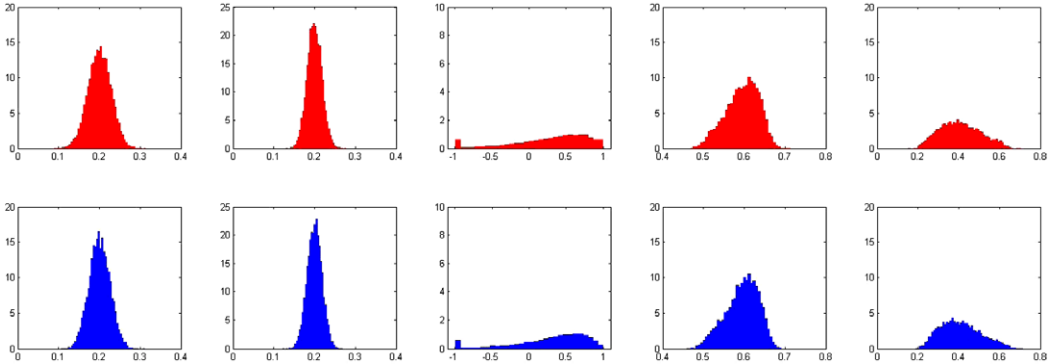
Similar to the previous subsection, we now describe the additional quantities needed to obtain the asymptotic distributions of the Wald statistics in the threshold-crossing model example. The derivations can similarly be found in Online Appendix B.

Recalling the function λ is defined in the Appendix, 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 $\Sigma(\pi; \gamma_0) =$



Asymptotic (bottom row) and finite-sample (top row, $n = 1000$) 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 2. Threshold crossing model parameter estimator densities when $b = \sqrt{n}0.1$.



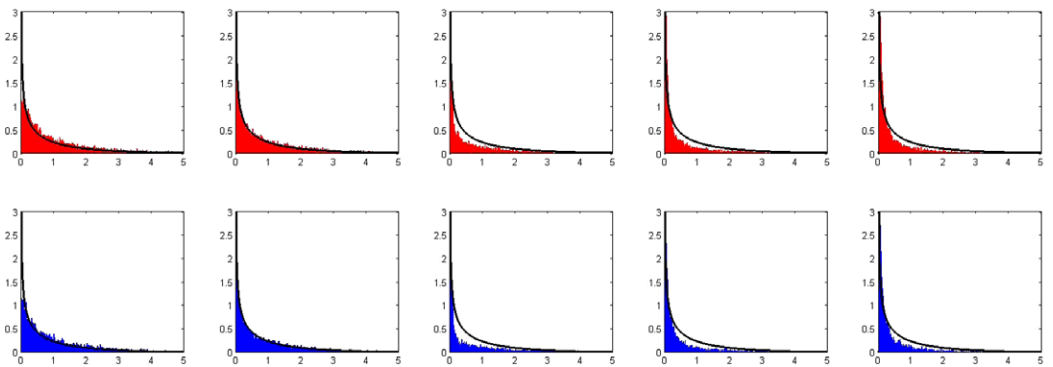
Asymptotic (bottom row) and finite-sample (top row, $n = 1000$) densities of the estimators of β, ζ, π_3, π_1 and π_2 (left-to-right) in the threshold-crossing model when $\zeta = 0.2$ and $\pi = (0.6, 0.4, 0.4)$.

FIGURE 3. Threshold crossing model parameter estimator densities when $b = \sqrt{n}0.2$.

$V^{-1}(\psi_0, \pi; \gamma_0)$, where

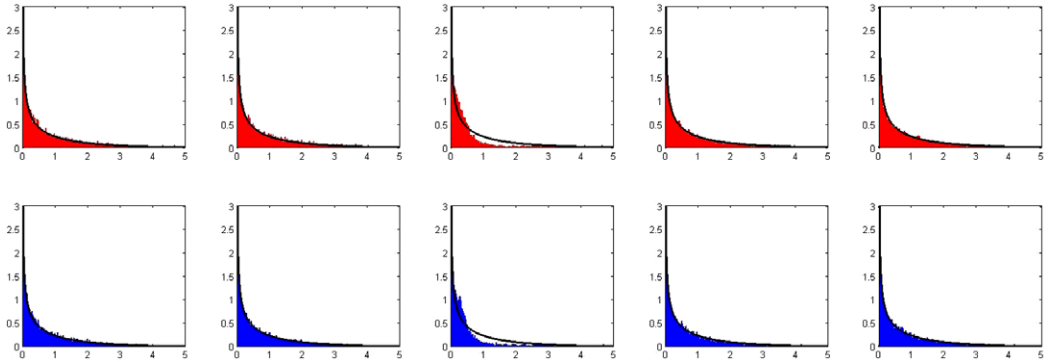
$$V(\psi_0, \pi; \gamma_0) = \sum_{y,d,z=0,1} \frac{\phi_{z,0}}{p_{yd,z}(\theta_0)} D_\theta p^\dagger_{yd,z}(\psi_0, \pi) D_\theta p^\dagger_{yd,z}(\psi_0, \pi)'.$$

Similar 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 4–6 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



Asymptotic (bottom row) and finite-sample (top row, $n = 1000$) 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^2_1 density overlay (bold line).

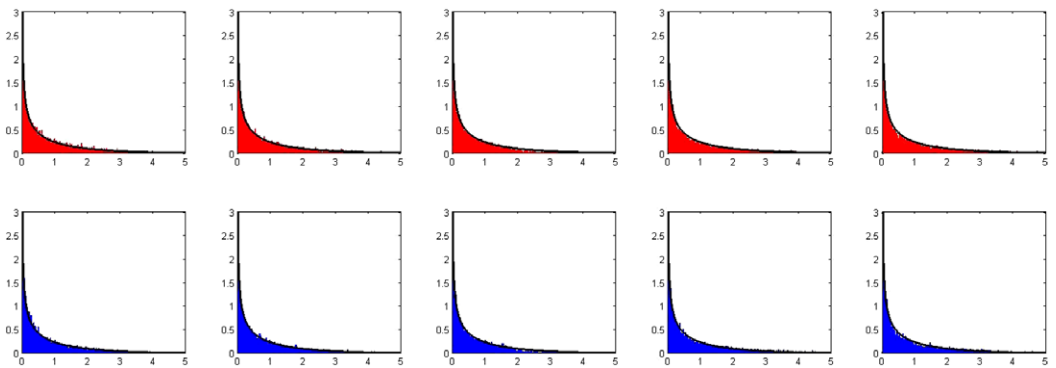
FIGURE 4. Wald statistic densities for the threshold crossing model when $b = 0$.



Asymptotic (bottom row) and finite-sample (top row, $n = 1000$) 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^2_1 density overlay (bold line).

FIGURE 5. Wald statistic densities for the threshold crossing model when $b = \sqrt{n}0.1$.

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 the top row and the densities of $\lambda(\pi^*_{0,b}; \gamma_0, b)$ are given in the bottom row. In addition, the solid black line graphs the density function of a χ^2_1 distribution for comparison. We look at identical true parameter values as in the previous subsection. Figures 4–6 show similar features to the corresponding figures for the estimators (Figures 1–3): (i) the distributions of the Wald statistics can depart significantly from the usual asymptotic χ^2_1 approximations in the presence of weak/nonidentification; (ii) as β grows larger, the distributions become approximately χ^2_1; and (iii) the new asymptotic distributional approximation perform very well, especially compared to the usual χ^2_1 approximation when β is small. One interesting additional feature to note is that, although the distributions of the parameter estimates when $\beta = 0.2$ in Figure 3 appear highly non-



Asymptotic (bottom row) and finite-sample (top row, $n = 1000$) 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^2_1 density overlay (bold line).

FIGURE 6. Wald statistic densities for the threshold crossing model when $b = \sqrt{n}0.2$.

Gaussian (especially for $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_3$), the corresponding distributions in Figure 6 look well approximated by the χ^2_1 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 (Forthcoming) 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.[28]

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 $\hat{I}^a_n(\hat{b}_n, \hat{\pi}_n)$ 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 7. 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 an 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. Lochner and Moretti (2004) studied whether increasing one's education level tends to reduce their engagement in criminal behavior and whether a policy aimed at doing so would be cost-effective relative to other crime prevention policies. These questions relate to measuring the social returns to education that incorporate spill-over effects or externalities. As part of their analysis, Lochner and Moretti (2004) measured the effect of schooling on incarceration using Census data. Since schooling is likely to be correlated with some of the nonmeasurable factors that

[28] 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 $\chi^2_5(0.95)^{-1}$.

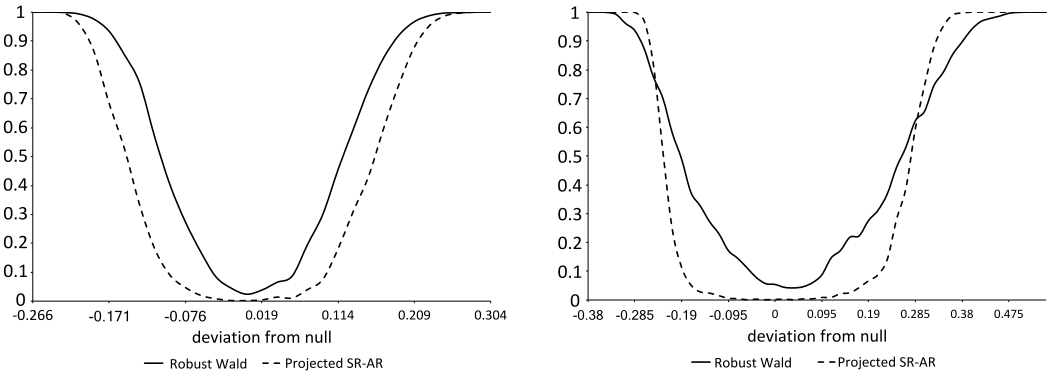


FIGURE 7. Power curves for testing $\boldsymbol{\pi}_2$ in the threshold crossing model.

determine incarceration, they use differences in state compulsory attendance laws to instrument for endogenous schooling. They find that schooling significantly reduces the probability of incarceration. For this application, we use the same US Census data (Lochner and Moretti's, 2004 "inmates" data). All data and descriptions thereof are freely available on Enrico Moretti's website (http://eml.berkeley.edu/~moretti/).

Of the many sets of variables examined by these authors, one fits particularly neatly into the following threshold crossing model of a triangular system (Example 2.3) we examine in detail in this paper:

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

where $C(\varepsilon, v; \boldsymbol{\pi}_3) = \varepsilon v/[1 - \boldsymbol{\pi}_3(1-\varepsilon)(1-v)]$ denotes the Ali–Mikhail–Haq copula. 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"). In this model, $\boldsymbol{\pi}_2 \equiv \boldsymbol{\pi}_1 + \tilde{\boldsymbol{\pi}}_2$ captures the counterfactual probability that an individual would be incarcerated had they dropped out of high school, $\boldsymbol{\pi}_1$ is the counterfactual probability that an individual would be incarcerated had they not dropped out, β measures how compulsory schooling laws affect the decision to drop out from high school, $\boldsymbol{\zeta}$ is equal to the probability of dropping out for an individual attending a school without a compulsory schooling law and $\boldsymbol{\pi}_3$ measures the dependence between the unobserved components driving the high school drop out decision and whether an individual becomes incarcerated.

A weak identification problem arises in the context of this application when the compulsory schooling laws have little (or no) effect on whether students drop out from high school, relative to the variability of the underlying data. The weakness of compulsory schooling laws as IVs for educational attainment decisions has been widely documented in previous literature (e.g., Staiger and Stock (1997)). When weak identification is present in this model, the usual large sample normal and chi-squared distributions

fail to produce good approximations to the sampling distributions of parameter estimators and Wald statistics, respectively (see Sections 7.1–7.2). This is analogous to a weak instruments problem in the standard linear IV model. However, the nonlinearity of the threshold crossing model complicates the analysis by making it difficult to determine exactly which (functions of) parameters are weakly versus strongly identified. This is precisely what our reparameterization method can be used to determine.

We focus on the subpopulation of black individuals as the response of black individuals to schooling may be different than that of white individuals. Lochner and Moretti (2004) also provide separate analyses for white versus 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: $\hat{\beta}_n = -0.0137$, $\hat{\boldsymbol{\zeta}}_n = 0.3060$, $\hat{\boldsymbol{\pi}}_{1,n} = 0.0260$, $\hat{\boldsymbol{\pi}}_{2,n} = 0.0782$, 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. In other words, the data indicate that although there is evidence that the presence of a compulsory schooling law decreases the probability that a student drops out, this effect is weak enough to produce sampling distributions for parameter estimators and test statistics that are not well approximated by the usual normal and chi-squared approximations. The nonrobust inference methods used in the original study by Lochner and Moretti (2004) rely upon the usual large sample normal and chi-squared approximations. In weakly identified scenarios, these types of approximations are likely to produce, for example, significance tests with actual size in exceedance of their nominal level and confidence intervals that undercover.

We perform robust Wald inference for the main parameter of interest $\boldsymbol{\pi}_2$. Rather than comparing the Wald statistic for this parameter to the upper quantiles of a chi-squared distribution (with one degree of freedom), we approximate the null distribution by a stochastic process evaluated at the minimizer of a noncentral chi-squared process to form a confidence interval. General expressions for these processes are given in Theorem 4.1(i) and Proposition 5.1(i). Expressions for the quantities entering these processes that are specific to the threshold-crossing model are given in Sections 7.1–7.2. Though their expressions are involved, these processes and the resulting distributional approximations are straightforward to simulate. By using this more accurate nonstandard approximation to the null distribution of the Wald statistic, we produce a confidence interval with true coverage much closer to the nominal level than would obtain under the usual chi-squared distributional approximation.

Though they are straightforward to compute via simulation, the upper quantiles of the nonstandard distributional approximations depend upon nuisance parameters ϕ_1, β, ζ, b, and π. To implement our robust inference procedures, we must determine the values of these nuisance parameters at which to simulate the distributions. The nuisance parameters $\phi_1 = P(Z_i = 1)$, β and $\zeta = (\boldsymbol{\zeta}, \boldsymbol{\pi}_1 - C(\boldsymbol{\pi}_1, \boldsymbol{\zeta}; \boldsymbol{\pi}_3), C(\boldsymbol{\pi}_2, \boldsymbol{\zeta}; \boldsymbol{\pi}_3))$ are consistently estimable via ML estimation and we evaluate the nonstandard distributions at

consistent estimates for these parameters. On the other hand, the nuisance parameters b and π are not consistently estimable. For these parameters, we compute the 0.5-level confidence set $\hat{I}_n^{0.5} = \{b \in \mathbb{R} : (\sqrt{n}\hat{\beta}_n - b)'\hat{\Sigma}_{\beta\beta,n}^{-1}(\sqrt{n}\hat{\beta}_n - b) \leq \chi_1^2(0.5)^{-1}\} \times \mathbb{R}$, where $\hat{\Sigma}_{\beta\beta,n}$ is $\sqrt{n}$ times the usual standard error for the ML estimator $\hat{\beta}_n$ and $\chi_1^2(0.5)^{-1}$ is the 50th percentile of a chi-squared distribution with one degree of freedom. This set provides us with the range of identification-strength parameters b that the data indicate as the most relevant to this particular application. We find the largest $1-\alpha$ quantile of the nonstandard distribution across the confidence set $\hat{I}_n^{0.5}$. Finally, to correct for the fact that (b, π) may not lie in the confidence set, we add a small constant to this largest quantile. This constant is determined by jointly simulating the nonstandard null asymptotic distribution of the Wald statistic and $\hat{\beta}_n$ according to Steps 1 and 2 in Section 6.2. The sum of this constant and largest quantile is equal to the identification-robust adjusted-Bonferroni CV. The robust Wald test rejects when the standard Wald statistic exceeds this CV.

For a test of nominal size $\alpha = 0.05$, the adjusted-Bonferroni CV just described is $c_{1-\alpha,n}^{\mathrm{AB}} \approx 11.5$.[29] 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 $\boldsymbol{\pi}_2$ (lb^* must be strictly greater than zero to satisfy the parameter space conditions in Online Appendix B). For comparison, the standard Wald confidence interval that is not robust to weak identification is equal to $[0.036, 0.12]$. It is interesting to note that, in contrast to the standard confidence interval, the robust confidence interval implies that we fail to reject *any* small value of the counterfactual probability. That is, no matter how close lb^* is to zero, we cannot reject the null hypothesis $H_0 : \boldsymbol{\pi}_2 = lb^*$.

Appendix: Assumptions in Andrews and Cheng (2012)

For the reader's convenience, we restate the assumptions of AC12 that appear in the theorems in Sections 4–6 of the current paper. At the end of this subsection, we also restate the expression for the limit distribution of the Wald statistic of AC14 that is used in Sections 5–6. "With respect to" is abbreviated as "w.r.t."

Recall $\bar{h}(\theta) \equiv (\beta, h(\mu))$. As noted in footnotes 15 and 18, the *true* parameter spaces for $\boldsymbol{\theta}$ and θ differ from the parameter spaces $\boldsymbol{\Theta}$ and Θ. Specifically, let $\boldsymbol{\Theta}^*$ denote the true parameter space for $\boldsymbol{\theta}$ and $\Theta^* \equiv \bar{h}^{-1}(\boldsymbol{\Theta}^*)$ denote the true parameter space for θ. Lemma 4.1(i) implies the bijectivity of $\bar{h} : \Theta^* \to \boldsymbol{\Theta}^*$, since we assume that the true parameter space is contained in the optimization parameter space (see Assumption B1 below). We assume that $\boldsymbol{\Theta}^*$ is a compact subset of $\mathbb{R}^{d_\theta}$. All statements made throughout the main text should technically be written with $\boldsymbol{\Theta}^*$ and Θ^* replacing $\boldsymbol{\Theta}$ and Θ in the definitions of $\boldsymbol{\Gamma}$ and Γ.

Define $\Theta_\delta^* \equiv \{\theta \in \Theta^* : \|\beta\| < \delta\}$, where Θ^* is the true parameter space for θ.

Assumption B1 (AC12). (i) $\mathrm{int}(\boldsymbol{\Theta}) \supset \boldsymbol{\Theta}^*$. (ii) *For some* $\delta > 0$, $\boldsymbol{\Theta} \supset \{\beta \in \mathbb{R}^{d_\beta} : \|\beta\| < \delta\} \times \mathcal{Z}^0 \times \Pi \supset \Theta_\delta^*$ *for some nonempty open set* $\mathcal{Z}^0 \subset \mathbb{R}^{d_\zeta}$. (iii) Π *is compact.*

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

ASSUMPTION B2 (AC12). (i) Γ *is compact and* $\Gamma = \{\gamma = (\theta, \phi) : \theta \in \Theta^*, \phi \in \Phi^*(\theta)\}$. (ii) *For some* $\delta > 0$, $\gamma = (\beta, \zeta, \pi, \phi) \in \Gamma$ *with* $0 \leq \|\beta\| < \delta$ *implies that* $\tilde{\gamma} = (\tilde{\beta}, \zeta, \pi, \phi) \in \Gamma$ *for all* $\tilde{\beta} \in \mathbb{R}^{d_\beta}$ *with* $0 \leq \|\tilde{\beta}\| < \delta$. (iii) *For* $\delta > 0$ *as in* (ii), $\exists \gamma = (\beta, \zeta, \pi, \phi) \in \Gamma$ *with* $0 < \|\beta\| < \delta$.

ASSUMPTION B3. (i) *For some nonstochastic real-valued function* $Q(\theta; \gamma_0)$ *on* $\Theta \times \Gamma$, $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta; \gamma_0)| \to_p 0$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0)$ $\forall \gamma_0 \in \Gamma$. (ii) *When* $\beta_0 = 0$, *for every neighborhood* $\Psi_0 (\subset \mathbb{R}^{d_\psi})$ *of* $\psi_0 = (\beta_0, \zeta_0)$, $\inf_{\pi \in \Pi} (\inf_{\psi \in \Psi(\pi)/\Psi_0} Q(\psi, \pi; \gamma_0) - Q(\psi_0, \pi; \gamma_0)) > 0$ $\forall \gamma_0 = (\psi_0, \pi_0, \phi_0) \in \Gamma$. (iii) *When* $\beta_0 \neq 0$, *for every neighborhood* $\Theta_0 (\subset \Theta)$ *of* $\theta_0 = (\beta_0, \zeta_0, \pi_0)$, $\inf_{\theta \in \Theta/\Theta_0} Q(\theta; \gamma_0) - Q(\theta_0; \gamma_0) > 0$ $\forall \gamma_0 = (\theta_0, \phi_0) \in \Gamma$.

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

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

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

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

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

ASSUMPTION C2. (i) $D_\psi Q_n(\theta)$ *takes the form* $D_\psi Q_n(\theta) = n^{-1} \sum_{i=1}^n m(W_i, \theta)$ *for some function* $m(W_i, \theta) \in \mathbb{R}^{d_\psi}$ $\forall \theta \in \Theta_\delta$, *for any true parameter* $\gamma^* \in \Gamma$. (ii) $E_{\gamma^*} m(W_i, \psi^*, \pi) = 0$ $\forall \pi \in \Pi$, $\forall i \geq 1$ *when the true parameter is* γ^* $\forall \gamma^* = (\psi^*, \pi^*, \phi^*) \in \Gamma$ *with* $\beta^* = 0$.

Define an empirical process $\{G_n(\pi) : \pi \in \Pi\}$ by

$$G_n(\pi) = n^{-1/2} \sum_{i=1}^n \big(m(W_i, \psi_{0,n}, \pi) - E_{\gamma_n} m(W_i, \psi_{0,n}, \pi)\big).$$

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

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

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

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

The domain of the function $K_n(\theta; \gamma^*)$ is $\Theta_\delta \times \Gamma_0$, where $\Gamma_0 = \{\gamma_a = (a\beta, \zeta, \pi, \phi) \in \Gamma : \gamma = (\beta, \zeta, \pi, \phi) \in \Gamma$ with $\|\beta\| < \delta$ and $a \in [0, 1]\}$, and $\delta > 0$ is as in Assumption B2(ii).

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

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

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

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* $\Pr_{\gamma_0}(A(\gamma_0, b)) = 1$ *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* $\|b\| < \infty$.

Define a nonstochastic function $\{\eta(\pi; \gamma_0, \omega_0) : \pi \in \Pi\}$ by

$$\eta(\pi; \gamma_0, \omega_0) \equiv -\frac{1}{2}\omega_0' K(\pi; \gamma_0)' H^{-1}(\pi; \gamma_0) K(\pi; \gamma_0)\omega_0.$$

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

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

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

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

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

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

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

ASSUMPTION D2. *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $J_n = B^{-1}(\beta_n)D^2Q_n(\theta_n)\times B^{-1}(\beta_n) \to_p J(\gamma_0) \in \mathbb{R}^{d_\theta \times d_\theta}$, *where* $J(\gamma_0)$ *is nonsingular and symmetric.*

ASSUMPTION D3. (i) *Under* $\{\gamma_n\} \in \Gamma(\gamma_0, \infty, \omega_0)$, $n^{1/2}B^{-1}(\beta_n)DQ_n(\theta_n) \to_d G^*(\gamma_0) \sim N(0_{d_\theta}, V(\gamma_0))$ *for some symmetric* $d_\theta \times d_\theta$ *matrix* $V(\gamma_0)$. (ii) $V(\gamma_0)$ *is positive definite* $\forall \gamma_0 \in \Gamma$.

Let

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

ASSUMPTION V1—Scalar β. (i) $\widehat{J}_n = \widehat{J}_n(\widehat{\theta}_n)$ *and* $\widehat{V}_n = \widehat{V}_n(\widehat{\theta}_n)$ *for some* (*stochastic*) $d_\theta \times d_\theta$ *matrix-valued functions* $\widehat{J}_n(\theta)$ *and* $\widehat{V}_n(\theta)$ *on* Θ *that satisfy* $\sup_{\theta\in\Theta} \|\widehat{J}_n(\theta) - J(\theta; \gamma_0)\| \to_p 0$ *and* $\sup_{\theta\in\Theta} \|\widehat{V}_n(\theta) - V(\theta; \gamma_0)\| \to_p 0$ *under* $\{\gamma_n\} \in \Gamma(\gamma_0, 0, b)$ *with* $\|b\| < \infty$. (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* $\beta_0 = 0$.

When β is a vector, we reparametrize β as $(\|\beta\|, \omega)$, where $\omega = \beta/\|\beta\|$ if $\beta \neq 0$ and by definition $\omega = 1_{d_\beta}/\|1_{d_\beta}\|$ with $1_{d_\beta} = (1, \ldots, 1) \in \mathbb{R}^{d_\beta}$ if $\beta = 0$. Correspondingly, θ is reparametrized as $\theta^+ = (\|\beta\|, \omega, \zeta, \pi)$. Let $\Theta^+ = \{\theta^+ : \theta^+ = (\|\beta\|, \beta/\|\beta\|, \zeta, \pi), \theta \in \Theta\}$. Let $\hat{\theta}_n^+$ and θ_0^+ be the counterparts of $\hat{\theta}_n$ and θ_0 after reparametrization. Let $J(\theta^+; \gamma_0)$ and $V(\theta^+; \gamma_0)$ denote some nonstochastic $d_\theta \times d_\theta$ matrix-valued functions such that $J(\theta^+; \gamma_0) = J(\gamma_0)$ and $V(\theta^+; \gamma_0) = V(\gamma_0)$. Let

$$\Sigma\big(\theta^+; \gamma_0\big) = J^{-1}\big(\theta^+; \gamma_0\big)V\big(\theta^+; \gamma_0\big)J^{-1}\big(\theta^+; \gamma_0\big),$$
$$\Sigma(\pi, \omega; \gamma_0) = \Sigma\big(\|\beta_0\|, \omega, \zeta_0, \pi; \gamma_0\big),$$

and let $\Sigma_{\beta\beta}(\pi, \omega; \gamma_0)$ denote the upper left $d_\beta \times d_\beta$ submatrix of $\Sigma(\pi, \omega; \gamma_0)$.

ASSUMPTION V1—Vector β. (i) $\hat{J}_n = \hat{J}_n(\hat{\theta}_n^+)$ *and* $\hat{V}_n = \hat{V}_n(\hat{\theta}_n^+)$ *for some* (*stochastic*) $d_\theta \times d_\theta$ *matrix-valued functions* $\hat{J}_n(\hat{\theta}_n^+)$ *and* $\hat{V}_n(\hat{\theta}_n^+)$ *on* Θ^+ *that satisfy* $\sup_{\theta^+\in\Theta^+} \|\hat{J}_n(\hat{\theta}_n^+) - J(\theta^+; \gamma_0)\| \to_p 0$ *and* $\sup_{\theta^+\in\Theta^+} \|\hat{V}_n(\hat{\theta}_n^+) - V(\theta^+; \gamma_0)\| \to_p 0$ *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^+ $\gamma_0 \in \Gamma$ *with* $\beta_0 = 0$. (iii) $\lambda_{\min}(\Sigma(\pi, \omega; \gamma_0)) > 0$ *and* $\lambda_{\max}(\Sigma(\pi, \omega; \gamma_0)) < \infty$ $\forall \pi \in \Pi$, $\forall \omega \in \mathbb{R}^{d_\beta}$ *with* $\|\omega\| = 1$, $\forall \gamma_0 \in \Gamma$ *with* $\beta_0 = 0$. (iv) $\Pr[\tau_\beta(\pi^*(\gamma_0, b), \gamma_0, b) = 0] = 0$ $\forall \gamma_0 \in \Gamma$ *with* $\beta_0 = 0$ *and* $\forall b$ *with* $\|b\| < \infty$.

ASSUMPTION V2. *Under* $\{\gamma_n\} \in \Gamma(0, \infty, \omega_0)$, $\hat{J}_n \to_p J(\gamma_0)$ *and* $\hat{V}_n \to_p V(\gamma_0)$.

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

Lastly, we restate the expression for the limit distribution of the Wald statistic. Define a stochastic process $\{\lambda(\pi; \gamma_0, b) : \pi \in \Pi\}$ by

$$\lambda(\pi; \gamma_0, b)$$
$$= \tau^A(\pi; \gamma_0, b)'\bar{B}(\pi; \gamma_0, b)\big(r_\theta^*(\psi_0, \pi)\bar{\Sigma}(\pi; \gamma_0, b)r_\theta^*(\psi_0, \pi)'\big)^{-1}\bar{B}(\pi; \gamma_0, b)\tau^A(\pi; \gamma_0, b),$$

where

$$\tau^A(\pi;\gamma_0,b)=\begin{pmatrix} r_\psi^*(\psi_0,\pi)\tau(\pi;\gamma_0,b) \\ A_2(\psi_0,\pi)\big(r(\psi_0,\pi)-r(\psi_0,\pi_0)\big)\end{pmatrix}\in\mathbb{R}^{d_r},$$

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

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

$$\omega^*(\pi;\gamma_0,b)=\frac{\tau_\beta(\pi;\gamma_0,b)}{\|\tau_\beta(\pi;\gamma_0,b)\|}.$$

References

Altonji, J. G., T. E. Elder, and C. R. Taber (2005), "An evaluation of instrumental variable strategies for estimating the effects of catholic schooling." *Journal of Human Resources*, 40 (4), 791–821. [1052]

Andrews, D. W. K. and X. Cheng (2012), "Estimation and inference with weak, semi-strong, and strong identificationith weak, semi-strong, and strong identification." *Econometrica*, 80, 2153–2211. [1021, 1022]

Andrews, D. W. K. and X. Cheng (2013), "Maximum likelihood estimation and uniform inference with sporadic identification failure." *Journal of Econometrics*, 173, 36–56. [1021]

Andrews, D. W. K. and X. Cheng (2014), "GMM estimation and uniform subvector inference with possible identification failure." *Econometric Theory*, 30, 287–333. [1021, 1023]

Andrews, D. W. K. and P. Guggenberger (Forthcoming), "Identification- and singularity-robust inference for moment condition models." *Quantitative Economics*. [1021, 1047, 1048, 1052, 1058]

Andrews, I. and A. Mikusheva (2016a), "Conditional inference with a functional nuisance parameter." *Econometrica*, 84, 1571–1612. [1021, 1047, 1048, 1052]

Andrews, I. and A. Mikusheva (2016b), "A geometric approach to weakly identified econometric models." *Econometrica*, 84, 1249–1264. [1021, 1047]

Antoine, B. and E. Renault (2009), "Efficient GMM with nearly-weak instruments." *Econometrics Journal*, 12, 135–171. [1023, 1040]

Antoine, B. and E. Renault (2012), "Efficient minimum distance estimation with multiple rates of convergence." *Journal of Econometrics*, 170, 350–367. [1023, 1040]

Arellano, M., L. P. Hansen, and E. Sentana (2012), "Underidentification?" *Journal of Econometrics*, 170, 256–290. [1022]

Choi, I. and P. C. B. Phillips (1992), "Asymptotic and finite sample distribution theory for IV estimators and tests in partially identified structural equations." *Journal of Econometrics*, 51, 113–150. [1022, 1023, 1034, 1040]

Cox, G. (2017), "Weak identification in a class of generically identified models with an application to factor models." Unpublished Manuscript, Department of Economics, Columbia University. [1022]

Dovonon, P. and E. Renault (2013), "Testing for common conditionally heteroskedastic factors." *Econometrica*, 81, 2561–2586. [1022]

Dunbar, G. R., A. Lewbel, and K. Pendaku (2013), "Children's resources in collective households: Identification, estimation, and an application to child poverty in Malawi." *American Economic Review*, 103, 438–471. [1027]

Escanciano, J. C. and L. Zhu (2013), "Set inferences and sensitivity analysis in semiparametric conditionally identified models." Unpublished Manuscript, Indiana University, Department of Economics. [1022]

Evans, W. N. and R. M. Schwab (1995), "Finishing high school and starting college: Do Catholic schools make a difference?" *The Quarterly Journal of Economics*, 110 (4), 941–974. [1052]

Goldman, D. P., J. Bhattacharya, D. F. Mccaffrey, N. Duan, A. A. Leibowitz, G. F. Joyce, and S. C. Morton (2001), "Effect of insurance on mortality in an HIV-positive population in care." *Journal of the American Statistical Association*, 96 (455), 883–894. [1052]

Han, S. (2010), "Identification and inference in a bivariate probit model with weak instruments." Unpublished Manuscript, Department of Economics, Yale University. [1019]

Han, S., and A. McCloskey (2019), "Supplement to 'Estimation and inference with a (nearly) singular Jacobian'." *Quantitative Economics Supplemental Material*, 10, https://doi.org/10.3982/QE989. [1023, 1024]

Han, S. and E. Vytlacil (2017), "Identification in a generalization of bivariate probit models with dummy endogenous regressors." *Journal of Econometrics*, 199, 63–73. [1026]

Heckman, J. J. (1979), "Sample selection bias as a specification error." *Econometrica*, 47, 153–161. [1025]

Heckman, J. J. and B. E. Honoré (1990), "The empirical content of the Roy model." *Econometrica*, 58, 1121–1149. [1026]

Jones, D. R. (2001), "A taxonomy of global optimization methods based on response surfaces." *Journal of Globabl Optimization*, 21, 345–383. [1051]

Jones, D. R., M. Schonlau, and W. J. Welch (1998), "Efficient global optimization of expensive black-box functions." *Journal of Global Optimization*, 13, 455–492. [1051]

Kleibergen, F. (2002), "Pivotal statistics for testing structural parameters in instrumental variables regression." *Econometrica*, 70, 1781–1803. [1021, 1052]

Kleibergen, F. (2005), "Testing parameters in GMM without assuming that they are identified." *Econometrica*, 73, 1103–1123. [1021]

Komunjer, I. and S. Ng (2011), "Dynamic identification of dynamic stochastic general equilibrium models." *Econometrica*, 79, 1995–2032. [1035]

Lee, L. F. and A. Chesher (1986), "Specification testing when score statistics are identically zero." *Journal of Econometrics*, 31, 121–149. [1022]

Lochner, L. and E. Moretti (2004), "The effect of education on crime: Evidence from prison inmates, arrests, and self-reports." *American Economic Review*, 94, 155–189. [1052, 1058, 1059, 1060]

McCloskey, A. (2017), "Bonferroni-based size-correction for nonstandard testing problems." *Journal of Econometrics*, 200, 17–35. [1023, 1050, 1051]

Moreira, M. J. (2003), "A conditional likelihood ratio test for structural models." *Econometrica*, 71, 1027–1048. [1021, 1052]

Phillips, P. C. B. (1989), "Partially identified econometric models." *Econometric Theory*, 5, 181–240. [1022, 1023, 1034, 1040]

Phillips, P. C. B. (2016), "Inference in near-singular regression." In *Essays in Honor of Aman Ullah* (R. C. H. G. Gonzalez-Rivera and T.-H. Lee, eds.), Advances in Econometrics, Vol. 36, 461–486, Emerald Group Publishing Limited. [1023, 1034, 1040]

Qu, Z. and D. Tkachenko (2012), "Identification and frequency domain quasi-maximum likelihood estimation of linearized dynamic stochastic general equilibrium models." *Quantitative Economics*, 3, 95–132. [1022, 1035]

Quarteroni, A., R. Sacco, and F. Saleri (2010), *Numerical Mathematics*, Vol. 37. Springer Science & Business Media. [1034]

Rhine, S. L., W. H. Greene, and M. Toussaint-Comeau (2006), "The importance of check-cashing businesses to the unbanked: Racial/ethnic differences." *Review of Economics and Statistics*, 88 (1), 146–157. [1052]

Rothenberg, T. J. (1971), "Identification in parametric models." *Econometrica*, 39, 577–591. [1020]

Sargan, J. D. (1983), "Identification and lack of identification." *Econometrica*, 51, 1605–1633. [1022, 1023, 1040]

Sims, C. A., J. H. Stock, and M. W. Watson (1990), "Inference in linear time series models with some unit roots." *Econometrica*, 58, 113–144. [1040]

Staiger, D. and J. H. Stock (1997), "Instrumental variables regression with weak instruments." *Econometrica*, 65, 557–586. [1020, 1052, 1059]

Stock, J. H. and J. H. Wright (2000), "GMM with weak identification." *Econometrica*, 68, 1055–1096. [1021]

Tommasi, D. and A. Wolf (2018), "Estimating household resource shares: A shrinkage approach." *Economics Letters*, 163, 75–78. [1027]

Co-editor Andres Santos handled this manuscript.

Manuscript received 27 September, 2017; final version accepted 27 November, 2018; available online 12 February, 2019.