Weak Identification in a Class of Generically Identified Models with an Application to Factor Models

Gregory Cox

January 21, 2017

Most Recent Version

Abstract

This paper provides new theorems for calculating the asymptotic distribution of extremum estimators along sequences of parameters that lead to an unidentified limit. These theorems are formulated for models that are doubly parameterized by structural and reduced form parameters. This paper applies these theorems to weak identification in factor models. Identification in factor models can be characterized in terms of a rank condition on the factor loadings. Weak identification in factor models is complicated by the fact that the boundary of the identified set may not be differentiable and the limit of the objective function may be degenerate. The new theorems are capable of handling these difficulties, yielding inference that is robust to failure of the rank condition. Explicit robust inference procedures are proposed for two example models: one model with one weak factor and one model with two factors that may be weak or entangled. This paper also provides an empirical application of inference for entangled factors in a model of parents investing in their children.

1 Introduction

Identification is an important condition for classical statistical theory. When identification fails, standard results on consistency and asymptotic normality of estimators, as well as valid inference using standard test statistics are no longer true. Furthermore, assumptions that ensure identification are often difficult to justify in empirical applications, and researchers may be interested to check the sensitivity of their results to an identification assumption. For these reasons, researchers benefit from robust inference procedures, which do not rely on identification assumptions to be valid.

This paper provides a robust inference procedure for a class of models that are generically identified. This means that almost everywhere in the parameter space, the parameters are identified, but there is a set of measure zero on which the parameters are not identified. Generally, whether or not the true parameter belongs to the set of measure zero can be formulated in terms of a rank condition.

In addition to being generically identified, the class consists of models that are doubly parameterized by structural and reduced form parameters. The reduced form parameters are always identified, while the structural parameters are only generically identified. The structural parameters can be divided into two types: ψ, which is always identified, and π, which may or may not be identified depending on the true value of ψ. The structure of identification is a characterization of the set of values of ψ that lead to π being unidentified. In doubly parameterized models, there exists a mapping from the structural to the reduced form parameters that encapsulates the structure of identification. For a fixed value of ψ, this mapping is invertible if and only if π is identified.

This paper advocates robust inference that proceeds in two steps. In the first step, a test determines whether or not the rank condition is close to failing. In the second step, a standard test statistic, either a Wald, a likelihood ratio, or a Lagrange multiplier test statistic, is compared to a critical value. If the rank condition is close to failing, the test statistic is compared to a robust critical value, while if the rank condition is not close to failing, the test statistic is compared to a standard chi-squared critical value. The robust critical value is chosen so that this procedure controls size whether or not identification holds.

Calculating the robust critical value involves characterizing the asymptotic distribution of the test statistic along drifting sequences of parameters that converge to points of identification failure. These sequences are weakly identified and induce nonstandard limit theory for estimators and test statistics. This paper provides new theorems for this nonstandard limit theory along weakly identified sequences of parameters within the class of generically identified, doubly parameterized models. The theorems assume regularity conditions on the reduced form, including a quadratic expansion on the objective function, and smoothness conditions on the mapping from structural to reduced form parameters. Under these assumptions, the theorems characterize the limiting distribution of an extremum estimator for the structural parameters.

These theorems are novel because of three technical difficulties that they can handle. First, the structure of identification is allowed to be very flexible, imposing only smoothness on the mapping from structural to reduced form parameters, rather than a particular form. This more flexible structure of identification is handled by taking a Taylor series expansion of the mapping that is globally valid over the identified set, out to the required order.

Second, the identified set for π, when π is not identified, may still be partially identified by bounds that are allowed to depend on the other parameters, ψ, in a possibly nondifferentiable way. This is handled by defining an auxiliary estimator, whose identified set does not depend on the other parameters, but drifts with the sample size. The discrepancy between the distribution of the auxiliary estimator and the distribution of the original estimator can be squeezed between the fixed, nondifferentiable boundary and the drifting boundary, showing that it is negligible.

Third, an important step in the proof of the theorems is characterizing the limit of the objective function as a stochastic process over the unidentified π parameters. Theorem 1 requires that this limit has a unique minimum over π almost surely. However, Theorem 2 allows for the possibility that this limit is degenerate and does not have a unique minimum. Theorem 2 handles this degeneracy by restandardizing the objective function to a higher rate so that higher-order terms that are nondegenerate become relevant in the limit.[1]

These theorems provide nonstandard limit theory for extremum estimators along weak sequences of parameters in the class of generically identified models. This limit theory extends to test statistics by the continuous mapping theorem. Robust critical values can be calculated by taking the supremum over quantiles of the asymptotic distributions for a comprehensive class of sequences. These robust critical values control asymptotic rejection probabilities whether or not the true value of the parameter is identified, providing a valid robust inference procedure.

The robust inference procedure presented in this paper is related to other robust inference procedures that have been proposed in the literature. Andrews and Cheng (2012, 2013, 2014) present a related procedure for robust inference. Their procedure requires a particular form for the structure of identification, which is that the magnitude of a parameter, β, determines the identification status of π. For some models, no reparameterization exists that satisfies this structure, including Examples 1 and 2 in this paper. Also, Andrews and Cheng (2012, 2013, 2014) do not allow for the boundary of the identified set for π to depend on the other parameters, ψ, nor do they allow for the limit of the objective function as a stochastic process over π to be degenerate.[2] Also note that there are models that the Andrews and Cheng approach covers that are not covered by this paper. In particular, models that do not have identified reduced form parameters do not fit into the class of models that this paper considers, but are covered by Andrews and Cheng (2012, 2013, 2014).

For a given model, finding a reparameterization that fits the structure of Andrews and Cheng (2012) may be difficult. Han and McCloskey (2016) provide a systematic reparameterization procedure that leads to a reparameterized model that satisfies the structure of Andrews and Cheng (2012). This reparameterization procedure may also be helpful for

[1] This method is similar to Sargan (1983) and Cho and White (2007), in which higher-order expansions are necessary for limit theory. This paper is different from Sargan (1983) in that the parameters may not be identified of any order. Instead, the higher-order terms are used for the limit theory. Cho and White (2007) provide limit theory for mixture models, while this paper provides limit theory for a general class of models.

[2] In Andrews and Cheng (2012), dependence of the boundary of the identified set on ψ is excluded by Assumption B1, and degeneracy in the limit of the objective function is excluded by Assumption C6.

satisfying the structure in this paper, specified in Section 3.

McCloskey (2012) advocates a robust inference procedure that takes the supremum over quantiles of the asymptotic distributions for a subset of the sequences, rather than a comprehensive class of sequences. The subset is determined in a data dependent way and size is controlled by a Bonferroni bound. This is an appealing alternative for translating the limit theory into a robust inference procedure.

I. Andrews and Mikusheva (2016a) provide a geometric approach to robust inference in minimum distance models. Their method considers the null hypothesis as a manifold in the reduced form parameter space and bounds a minimum distance test statistic using a bound on the curvature of the null hypothesis. This bound holds for an asymptotically normal estimator for the reduced form parameters. In contrast, this paper tests the hypothesis in the structural parameter space using standard test statistics and characterizing nonstandard limit theory under weak identification.

Chen, Christensen, O'Hara, and Tamer (2016) advocate a procedure for robust inference based on Markov Chain Monte Carlo simulations to construct a confidence set for the identified set. They assume a quadratic expansion on the reduced form parameters that is closely related to Assumption QE in this paper. They prove Bernstein-von Mises type theorems for the posterior distribution of test statistics, while this paper characterizes the asymptotic distributions of test statistics along weakly identified sequences of parameters.

There is also a very wide literature on robust inference in models defined by moment equalities. These approaches are valid alternatives, but come with some drawbacks. First, many of these approaches rely on special statistics for robust inference.[3] In contrast, this paper uses standard test statistics that are commonly used for standard inference and calculates new critical values which are robust to weak identification. This implies that under identification, the robust inference procedure numerically agrees with standard inference with probability approaching 1. Second, many of these approaches work well for full vector inference, but rely on projection or plug-in methods for subvector inference.[4] The projection methods can be very conservative and the plug-in methods require the nuisance parameters to be identified. However, subvector inference in this paper follows naturally from the limit theory for standard test statistics of subvector hypotheses. Finally, the moment equality approach defines weak identification according to the rate of convergence of the moments, or the Jacobian.[5] In contrast, this paper defines weak identification according to the invertibility of a mapping from structural to reduced form parameters. The invertibility definition

[3] Examples include Kleibergen (2005, 2007), I. Andrews (2016), Andrews and Guggenberger (2015, 2016), and I. Andrews and Mikusheva (2016b).

[4] Examples include Dufour and Taamouti (2005), Guggenberger and Smith (2005), Otsu (2006), and Chaudhuri and Zivot (2011).

[5] See Assumption C in Stock and Wright (2000).

is more naturally related to identification, defined as the existence of a mapping from the distribution of the data to the parameters. Furthermore, the reduced rank of the Jacobian is the mechanism through which weak identification affects the first order asymptotics. The invertibility definition allows a more full characterization of the effect of weak identification on the asymptotics, to an arbitrary order. For these reasons, this paper advocates using the invertibility definition of weak identification and robust inference using the robust critical values in this paper.

Factor models have been used for a variety of applications in economics. One way that factor models are used is by specifying structure on the factors and the errors and estimating the distribution of the factors by (quasi) maximum likelihood.[6] This type of factor model has been studied by Anderson and Rubin (1956), Lawley and Maxwell (1971), and in a high dimensional context by Bai and Li (2012, 2016).

Another way that factor models are used is to define the factors to be those random variables that explain a maximal amount of the covariation between the observed variables. These models tend to be estimated by principal components in a high dimensional context.[7] Within this framework, Onatski (2012) has addressed the question of the asymptotic distribution of the principal components estimator with weak factors. However, weak factors in a high dimensional context are very different. In particular, the definition of weak factors means that the strength of the signal has the same order as the noise. Weak factors in a high dimensional context have the same strength as strong factors in a low dimensional context. This paper considers weak factors in a low dimensional context that, in additional to not dominating the noise, may also fail to be identified.

Also, Kleibergen (2009) considers tests of risk premia in linear factor models. Although the risk premia may be weakly identified, the distribution of the factors is always strongly identified because that paper considers the case where the factors are observed.

The identification status of factor models estimated by maximum likelihood can be characterized by a rank condition on the factor loadings. When the rank condition fails, the distribution of the factors is not identified. This occurs when a factor is weak, the number of true factors is less than the specified number, or the factors are entangled, which means that observed variables depend on the same linear combination of the factors. Simulations by Briggs and MacCallum (2003) and Ximénez (2006, 2007, 2009, 2015) have shown that standard estimation methods have difficulty detecting weak factors, as defined by the rank condition on the factor loadings.

[6] Factor models have been used in this way to study school quality (see Black and Smith (2006) and Bernal, Mittag, and Qureshi (2016)), personality psychology in economics (see Almlund, Duckworth, Heckman, and Kautz (2011)), and parental investments in the skills of their children (see Cunha, Heckman, and Schennach (2010)), among other applications.

[7] See Bai and Ng (2002).

The previous literature on identification in factor models provides sufficient conditions for identification generically over the parameter space.[8] This paper provides a robust inference procedure for the distribution of the factors, as well as the factor loadings, with respect to failure of the rank condition. This is done by applying the general two step method. This paper divides sequences of parameters converging to points of rank condition failure into classes and applies Theorems 1 and 2 to characterize the asymptotic distributions of test statistics within each class. Robust critical values are calculated by taking the supremum over the quantiles of these asymptotic distributions. The resulting robust inference procedure is uniformly valid over failure of the rank condition. To my knowledge, this is the only paper that provides robust inference in factor models with respect to failure of identification.

The asymptotic distributions calculated for factor models tend to have mass points at the endpoints of the identified set. These mass points correspond to Heywood cases, when a variance parameter is estimated to be zero. A long-standing puzzle in factor models is explaining why Heywood cases occur so often in finite samples.[9] The limit theory in this paper provides an explanation by characterizing asymptotic distributions for which Heywood cases occur with positive probability in the limit.

This paper applies the robust inference approach to two example factor models. In the first example, there is only one factor that may be unidentified. The rank condition reduces to the number of nonzero factor loadings. The factor is strong if three factor loadings are nonzero. The factor is unidentified if one or more factor loadings are zero. If only one factor loading is nonzero, then the true model essentially has zero factors. This is a simple example in which the technical difficulties of a nondifferentiable boundary of the identified set and degeneracy in the limit of the objective function are present. Theorems 1 and 2 are required to characterize the asymptotic distribution of test statistics along sequences of parameters converging to points of rank condition failure.

In the second example, there are two factors, one or both of which may not be identified. The rank condition allows for three ways that the factors may not be identified. First, one of the observed variables may be irrelevant, in the sense that both factor loadings associated with that observed variable are zero. Second, the second factor may be weak in the sense that it does not have three factor loadings that are nonzero. This includes the case where only one factor loading is nonzero, so the model has only one factor. Third, the factors may be entangled in the sense that observed variables depend on the same linear combination of the two factors. This paper applies the general two step inference approach to this example, providing inference that is robust to rank condition failure.

This paper applies the second example to an empirical model of parents investing in

[8] For example, see Shapiro (1985) and Bekker and ten Berge (1997).

[9] See Heywood (1931) and the discussion of Heywood cases in Bollen (1989).

their children.[10] Cunha, Heckman, and Schennach (2010) estimate the production function of skills in children as a function of parental investments. The model assumes that a variety of observed variables of the home environment of the children can be summarized in two parental investment factors—investment in cognitive skills and investment in noncognitive skills. However, when they take the model to the data, they assume that there is only one type of parental investments out of a concern for identification failure. This assumption eliminates important questions about the relative effects of parental investment in cognitive versus noncognitive skills. For example investment in noncognitive skills may be more effective for developing skills at a different age than investment in cognitive skills, or there may be complementarities between investment in cognitive skills for one age and investment in noncognitive skills for another age. In an attempt to allow for these more nuanced questions, this paper takes the two parental investment model to the same data and performs robust inference, which does not require identification, on the distribution of the parental investment factors. I find that for one age category the factors are entangled and not identified, while for all the other age categories the factors are identified.

The rest of the paper proceeds as follows. Section 2 defines factor models, characterizes identification in factor models according to a rank condition, and describes an example with one factor. Section 3 defines a class of models and states theorems for characterizing asymptotic distributions of an extremum estimator along weakly identified sequences. Section 4 provides a procedure for robust inference in Example 1. Section 5 provides a procedure for robust inference in Example 2. Section 6 reports the results of robust inference for the distribution of parental investments in the cognitive and noncognitive skills of their children. Section 7 concludes. An appendix contains additional calculations and proofs. There are three additional documents, called Supplemental Materials 1, 2, and 3. Supplemental Materials 1 proves Theorems 1 and 2 for the general class of models and supplies additional limit theory for strong and semi-strong sequences. Supplemental Materials 2 provides the details for Example 1, a one factor model with a weak factor, stating and proving a theorem for robust inference. Supplemental Materials 3 provides the details for Example 2, a two factor model with possibly unidentified factors, stating and proving a theorem for robust inference.

2 Identification in Factor Models

This section discusses factor models which are a motivating example of the general theory in Section 3. Factor models postulate the existence of unobserved variables (factors) that

[10] Other papers in this literature that use factor models in this way are Cunha and Heckman (2008), Heckman, Pinto, and Savelyev (2013), Attanasio, Cattan, Fitzsimons, Meghir, and Rubio-Codina (2015), Attanasio, Meghir, and Nix (2015), Lekfuangfu (2015), Agostinelli and Wiswall (2016a, 2016b), and Pavan (forthcoming).

explain the covariation between observed variables. A factor model is defined by the equation:

$$X_i = \Lambda F_i + \epsilon_i, \tag{2.1}$$

where X_i is a p-vector of observed variables, F_i is an m-vector of unobserved factors, ϵ_i is a p-vector of unobserved error terms, and Λ is a $p \times m$ matrix of coefficients called factor loadings. Let the covariance matrix of the factors be given by Σ and the covariance matrix of the errors be given by a diagonal matrix, Φ. The parameters in Λ, Σ, and Φ can be grouped together in a vector of parameters, θ.

Factor models can be used in a variety of ways depending on how the factors are defined. This paper considers the case where the factors are assumed to be uncorrelated with the errors and a normalization is placed on the factor loadings. This setup gives the model a flexible measurement error interpretation, where the observed variables can be thought of as measurements or proxies of the factors.

Assumption Normalization.

$$\Lambda = \left[\begin{array}{c} I_m \\ \tilde{\Lambda} \end{array} \right].$$

Remarks:

1. Assumption Normalization specifies that the upper $m \times m$ block of the factor loadings is equal to the identity matrix. $\tilde{\Lambda}$ is a $p - m \times m$ matrix of parameters to be estimated.

2. Assumption Normalization solves an indeterminacy in the covariance equation,

$$\mathrm{Var}(X_i) = \Lambda \Sigma \Lambda' + \Phi.$$

For any invertible $m \times m$ matrix M, $\Lambda \Sigma \Lambda' = \Lambda M M^{-1} \Sigma M'^{-1} M' \Lambda' = \bar{\Lambda} \bar{\Sigma} \bar{\Lambda}'$, where $\bar{\Lambda} = \Lambda M$ and $\bar{\Sigma} = M^{-1} \Sigma M'^{-1}$. This shows that there are m^2 degrees of indeterminacy in the covariance equation, matching the m^2 number of restrictions imposed by the normalization.

3. Assumption Normalization can be interpreted as fixing a definition of the factors. First, the diagonal terms in the identity matrix specify the scale of the factors to be measured in terms of the units of the first m observed variables. Second, the off diagonal zeros in the identity matrix place restrictions on the covariation between the factors and the observed variables. For example, the first factor is defined to be that common component to the covariation in the data that is uncorrelated with observed variables numbered 2 through m.

4. In practice, the ordering of the variables is very important. The researcher should choose very carefully which observed variables are used for the normalization and which are free. For example, in the model of parental investments considered in Section 6, the observed variables include the number of books the child owns and the number of times the child goes to a museum. In this case, normalizing the parental investments in cognitive skills to be measured in terms of the number of books is appropriate because owning more books is not an investment in noncognitive skills. On the other hand, normalizing the parental investments in cognitive skills to be measured in terms of the number of trips to a museum would be inappropriate because a trip to a museum constitutes investment in both cognitive and noncognitive skills. This illustrates that the selection of the normalizing variables is important and the appropriateness of a given selection depends on the application.

Identification in this model is determined by covariance matrix matching, where the covariance of X_i is given by:

$$\mathrm{Var}(X_i) = \Omega(\theta) := \Lambda\Sigma\Lambda' + \Phi. \tag{2.2}$$

The left hand side is always identified while the right hand side is a nonlinear function of the parameters in the model. The model is identified if and only if this nonlinear function can be inverted, solving for the parameters.

There are considerable technical complications that arise when searching for conditions under which the nonlinear function can be inverted. This comes from the fact that inversion depends on the true value of the parameters. In response to this, the literature on identification in factor models has focused on verifying identification generically over the parameter space. For example, Shapiro (1985) states that

> "practitioners tend to disregard the problem, believing that for 'not too big' values of r the factor analysis model is 'usually' identified. This suggests an investigation of the factor analysis model from the *generic* point of view."[11]

He goes on to characterize identification generically by comparing the number of factors, m, to the number of observed variables, p. He conjectures that the Ledermann[12] bound,

$$\frac{2p + 1 - \sqrt{8p+1}}{2},$$

gives the maximum number of factors that are generically identified by p observed variables. This formula comes from counting the number of separate equations in (2.2) compared to

[11] Italics in the original. In that paper, r is the number of factors, equivalent to m in this paper.

[12] First calculated by Ledermann (1937).

the number of parameters. Shapiro proved that for m less than or equal to the Ledermann bound, the factor model is generically locally identified. He conjectured that for m strictly less than the Ledermann bound, the factor model is generically globally identified, which was proved by Bekker and ten Berge (1997).

While Shapiro considers only generic identification, proving the validity of standard estimation and inference procedures requires full identification. The problem is that loss of identification, even though it occurs on a set of measure zero in the parameter space, distorts the finite sample distributions of the estimator and test statistics whenever the truth is in a neighborhood of the loss of identification. This neighborhood may be shrinking with the sample size, but is not negligible in the finite sample. The simulation evidence at the end of this section shows that standard approaches are insufficient and an inference procedure that explicitly considers identification failure is needed.

In factor models, a rank condition can be placed on the factor loadings, which determines the identification status of the distribution of the factors. Anderson and Rubin (1956) provide one such rank condition that is necessary and sufficient for identification when the number of factors is one or two. This condition is sufficient if there are more factors, but not necessary. Appendix C provides recommendations for models with more than two factors.

Assumption Rank Condition. *Let Λ_{-j} denote the matrix formed by deleting the j^{th} row of Λ. If, for every $j = 1, ..., p$, the rows of Λ_{-j} can be rearranged into two matrices, both of full rank, m, then Λ satisfies the rank condition.*

Remarks:

1. This rank condition is on the factor loadings, a matrix of parameters, and is different from a rank condition on the Jacobian of a set of moments, which is typically given as a sufficient condition for local identification.[13]

2. The rank condition is an assumption that we do not want to make. It may be difficult to justify in a particular application, or a researcher may want to check the sensitivity of results to this assumption. This paper proceeds without this assumption by analyzing the consequences of rank condition failure.

3. The statement of the rank condition is enigmatic, but reduces to more intuitive conditions in the examples. Intuitively, the rank condition ensures that there is enough information about the factors in the covariance structure of the observed variables. This condition implies that for each factor, there are at least three observed variables with nonzero factor loadings, a necessary condition for identification.

[13] See Rothenberg (1971) for a rank condition on a Jacobian.

Example 1: One Factor

Consider the case where there is only one factor and three observed variables. In this case, the factor loadings matrix is given by:

$$\Lambda = \begin{bmatrix} 1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix},$$

where $\lambda_1 = 1$ represents the normalization. In this case, the rank condition reduces to: $\lambda_2 \neq 0$ and $\lambda_3 \neq 0$. The other parameters in the model are σ^2, the variance of the factor, and ϕ_1, ϕ_2, and ϕ_3, the variances of the errors.

In this example, we allow λ_2 and λ_3 to be close to or equal to zero. This means that there are two ways the rank condition, and consequentially identification of the parameters, can fail. The first way is that one of λ_2 or λ_3 is equal to zero while the other is nonzero. In this case, only two of the variables are related to the factor, which is not enough for identification. However, since there is some covariation, the factor is present and detectable by the data.

The second way the rank condition can fail is if both λ_2 and λ_3 are equal to zero. This is essentially a zero factor model because there is no factor that is common to the observed variables. In this sense, this example encompasses an unknown number of factors (either zero or one).

Simulations

This simulation section shows that standard inference procedures are insufficient in Example 1, and an inference procedure that explicitly considers identification failure is needed. The true values of the parameters are:

$$\sigma^2 = 1, \qquad \Phi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and the values of the factor loadings, λ_2 and λ_3, are allowed to vary. The sample size is $n = 500$, a typical sample size for the empirical application. With 10,000 simulations, I calculate the finite sample distribution of $\hat{\sigma}_n^2$, the Gaussian maximum likelihood estimator of the factor variance, σ^2, as well as two asymptotic approximations. Figure 1 shows the finite sample distribution (left column) for $\hat{\sigma}_n^2$ compared to the asymptotic normal approximation (center column) and a weak approximation (right column). The weak asymptotic approximation is

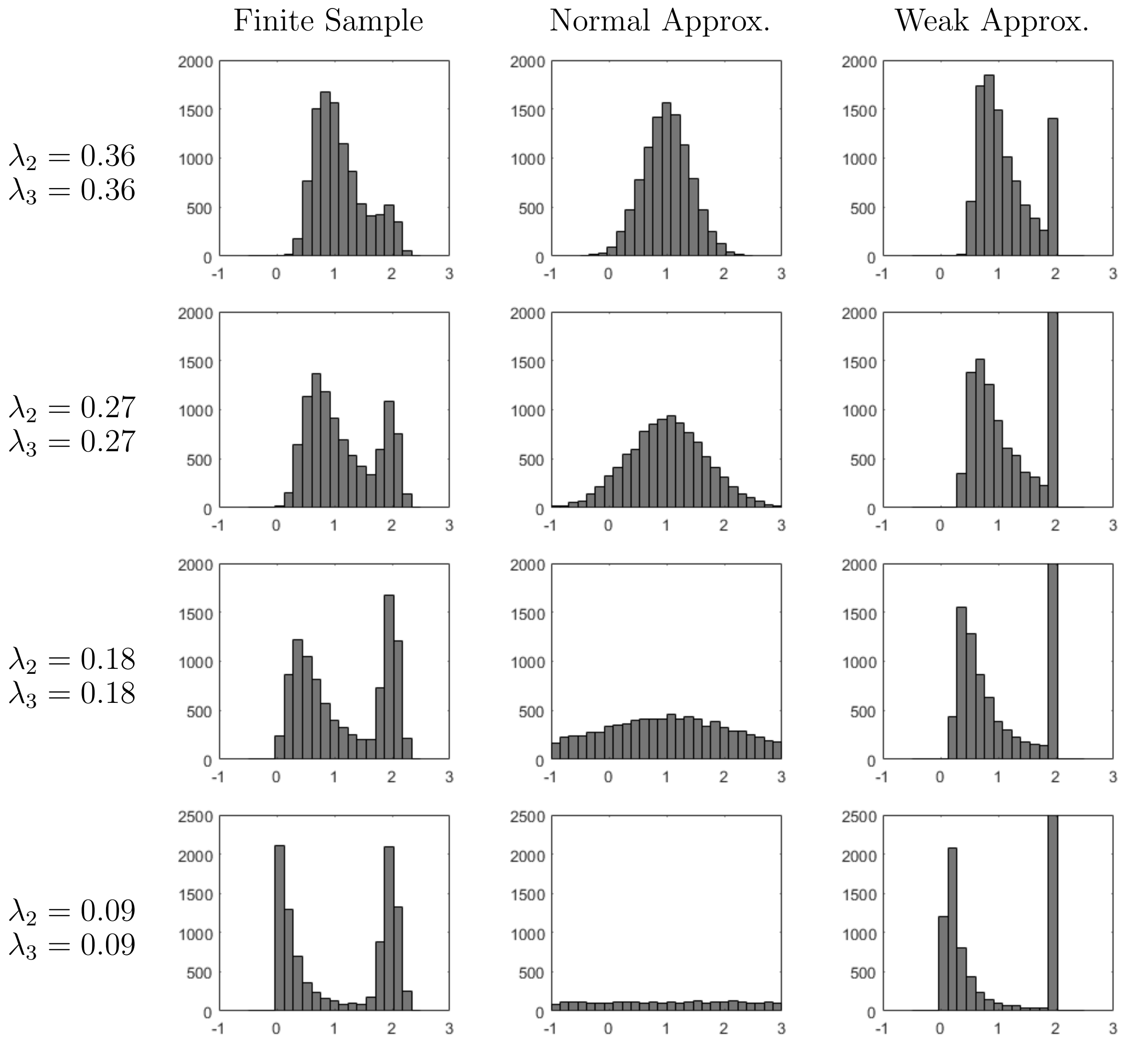

Figure 1: Finite sample ($n = 500$), normal approximation, and weak approximation densities of $\hat{\sigma}_n^2$ in Example 1 when the true value of σ^2 is 1.

derived from a sequence of true values of the parameters converging to a point where $\lambda_2 = 0$ and $\lambda_3 = 0$. For the simulations, the sequence is indexed by a local parameter that is given by $b = n^{1/2}\beta_1\beta_2 = n^{1/2}\lambda_2\lambda_3\sigma^4$. More details on the weak asymptotic approximation can be found in Section 4.1.

Also, Table 1 reports rejection rates of the null hypothesis, $H_0 : \sigma^2 = \sigma_0^2$ using the Wald test statistic, $W_n(\sigma_0^2)$, compared to standard chi-squared critical values.

Remarks on Figure 1 and Table 1:

1. The variance of the normal approximation diverges as λ_2 and λ_3 approach zero. This is because of the loss of identification of σ^2. The likelihood becomes flatter, corresponding to an increase in the asymptotic variance.

λ_2	λ_3	Rejection Rate
0.36	0.36	0.076
0.27	0.27	0.113
0.18	0.18	0.181
0.09	0.09	0.296
0	0	0.363

Table 1: Finite sample ($n = 500$) null rejection rates for a Wald test of the value of σ^2 in Example 1.

2. Both the shape and the dispersion of the finite sample distribution are poorly approximated by the shape and dispersion of the normal approximation. However, the shape and dispersion of the finite sample distribution are well approximated by the shape and dispersion of the weak approximation.

3. The quality of the normal approximation deteriorates as the values of λ_2 and λ_3 go to zero. However, the weak approximation retains it high quality as the values of λ_2 and λ_3 go to zero. In fact, the normal approximation is poor for very significant values of the factor loadings. Ximénez (2007) performs related simulations and notes that values of the factor loadings between 0.25 and 0.35 would not be considered "weak" by practical standards.[14]

   Weak asymptotics provides an explanation for the poor approximation for relatively large values of λ_2 and λ_3. There exist weak sequences of parameters that converge to points of rank failure at the $n^{-1/4}$ rate.[15] Compare this to the weak instrumental variables (IV) model with one endogenous variable and one instrument, in which all weak sequences converge to a point of rank failure at the $n^{-1/2}$ rate or faster.[16] This means that, compared to the weak IV model, the area of the parameter space that is affected by identification failure is much larger and shrinks much more slowly as the sample size increases. This explains why weak identification has an effect for relatively large values of the factor loadings. In this sense, the loss of identification in factor models is more significant than the corresponding loss of identification in weak IV models.

4. The finite sample distribution contains spikes near the left and right endpoints of the

[14] Briggs and MacCallum (2003) and Ximénez (2006, 2007, 2009, 2015) run simulations to determine the ability of various estimation methods to detect weak factors.

[15] For example, $\lambda_{1n} = \lambda_{2n} = n^{-1/4}$. Section 4.1 describes the different types of weak sequences in Example 1.

[16] See Staiger and Stock (1997).

support. These spikes are also present in the weak approximation, but missing in the normal approximation. These spikes correspond to Heywood cases that occur when one of the variance parameters, either the variance of the factor or one of the variances of the errors, is estimated to be zero.[17] One of the long-standing puzzles in factor models is explaining why Heywood cases occur so often in practice. Weak asymptotics provides one explanation, by characterizing asymptotic distributions for which Heywood cases occur with positive probability in the limit.

5. Looking now at Table 1, loss of identification leads to significant over-rejection when the rank condition is close to failing. This over-rejection is as high as 36% for $\lambda_2 = \lambda_3 = 0$ and persists as λ_2 and λ_3 get larger, having a non-negligible effect when the factor loadings are as large as $\lambda_2 = \lambda_3 = 0.36$.

Overall, these simulations show that under rank condition failure, the normal approximation is poor, the weak approximation is good, and the standard test based on the Wald test statistic significantly over-rejects.

3 Robust Inference for a General Class of Weakly Identified Models

This section provides a general approach and useful tools for robust inference. Section 3.1 defines a class of models that includes factor models. Sections 3.2 and 3.3 state theorems for characterizing asymptotic distributions along drifting sequences.

3.1 A Class of Models

This section defines a class of models for robust inference. The class can be described as doubly parameterized by structural and reduced form parameters. Thus, the model has two parameter spaces and two objective functions.

Definition Parameter Spaces.

(a) Let $\Theta \subset \bar{\Theta} \subset \mathbb{R}^{d_\theta}$.

(b) For each $\theta \in \bar{\Theta}$, let $\theta = (\psi, \pi)$ and $d_\theta = d_\psi + d_\pi$.

(c) Let $h : \bar{\Theta} \to \mathbb{R}^{d_h}$.

(d) Let $\delta : \bar{\Theta} \to \mathbb{R}^{d_\delta}$ be defined by: $\delta(\theta) = (\psi, h(\psi, \pi))$.

[17] See Heywood (1931) and the discussion of Heywood cases in Bollen (1989).

(e) Let $\Delta = \delta(\bar{\Theta})$, and let $\bar{\Delta} \supset \Delta$.

Remarks:

1. Part (a) defines the structural parameter space, Θ. Θ is equipped with all of the structure of a Euclidean space, including Lebesgue measure. θ is the parameter of interest. $\bar{\Theta}$ is a slight expansion of the parameter space.

2. Part (b) separates the parameters in θ into two types, ψ and π. The different roles of ψ and π are described later. d_ψ and d_π denote the lengths of the corresponding vectors. In general, d_X denotes the length of the vector X.

3. Part (c) defines $h(\psi, \pi)$, a mapping from structural parameters to new parameters, $h \in \mathbb{R}^{d_h}$.

4. Part (d) defines δ, a mapping from the structural parameters defined by the concatenation of the identity (for ψ) and h. δ are the reduced form parameters. Abusing notation, we use δ and h to denote both the mapping from the structural parameters and the reduced form parameters themselves.

5. Part (e) defines Δ and $\bar{\Delta}$, the reduced form parameter space and a slight expansion.

The reduced form parameter, $\delta = (\psi, h)$, is always identified. The parameter that may not be identified is π. If, for fixed ψ, $h(\psi, \pi)$ can be inverted for a value of π, then π is identified. However, if $h(\psi, \pi)$ cannot be inverted for π, then π is not identified. In this way, the value of ψ determines the identification status of π through the invertibility of $h(\psi, \pi)$.

Definition Objective Functions.

(a) Let $\mathbb{X}_n$ denote the data with sample size n.

(b) For each $\delta \in \bar{\Delta}$, let $Pr_{n,\delta}$ denote a distribution of $\mathbb{X}_n$.[18]

(c) For each $\delta \in \bar{\Delta}$, let $T_n(\delta)$ be a real valued random variable (a measurable function of $\mathbb{X}_n$).

(d) For each $\theta \in \bar{\Theta}$, let $Q_n(\theta) = T_n(\delta(\theta))$.

Remarks:

[18] The distribution of $\mathbb{X}_n$ may also be indexed by another (possibly infinite dimensional) nuisance parameter, suppressed for notational simplicity. For example, the error distribution is an infinite dimensional nuisance parameter that indexes factor models. This section requires that the parameter that determines the identification status of π is finite dimensional and belongs to ψ.

1. $Pr_{n,\delta}$ is sometimes written Pr_δ or Pr_θ, suppressing the dependence on the sample size. Pr_θ is short for $Pr_{\delta(\theta)}$.
2. $T_n(\delta)$ is the reduced form objective function, defined over the reduced form parameters. $Q_n(\theta)$ is the structural objective function, defined over the structural parameters.
3. This definition allows for (quasi) maximum likelihood, if $Q_n(\theta)$ is a likelihood that only depends on reduced form parameters, or minimum distance, in which $T_n(\delta)$ is a quadratic form in δ and depends on the data through a first stage estimator of the reduced form parameters.

The $h(\psi, \pi)$ function is the key function, providing the link between the structural and reduced form parameter spaces as well as the link between the structural and reduced form objective functions.

Example 1, Continued. Example 1 can be reparameterized into the setup of Definition Parameter Spaces and Definition Objective Functions. The reduced form parameters in this model are the elements of the covariance matrix of the observed variables, equal to

$$Cov(X_i) = \Omega(\theta) = \begin{bmatrix} \sigma^2 + \phi_1 & \lambda_2\sigma^2 & \lambda_3\sigma^2 \\ \lambda_2\sigma^2 & \lambda_2^2\sigma^2 + \phi_2 & \lambda_2\lambda_3\sigma^2 \\ \lambda_3\sigma^2 & \lambda_1\lambda_2\sigma^2 & \lambda_3^3\sigma^2 + \phi_3 \end{bmatrix}.$$

We can define a reparameterization by:

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \pi \end{pmatrix} = \begin{pmatrix} \lambda_2\sigma^2 \\ \lambda_3\sigma^2 \\ \sigma^2 + \phi_1 \\ \lambda_2^2\sigma^2 + \phi_2 \\ \lambda_3^2\sigma^2 + \phi_3 \\ \sigma^2 \end{pmatrix}.$$

Then, the covariance matrix becomes:

$$\Omega(\theta) = \begin{bmatrix} \zeta_1 & \beta_1 & \beta_2 \\ \beta_1 & \zeta_2 & h(\psi, \pi) \\ \beta_2 & h(\psi, \pi) & \zeta_3 \end{bmatrix},$$

where $\psi = (\beta_1, \beta_2, \zeta_1, \zeta_2, \zeta_3)$ and $h(\psi, \pi) = \beta_1\beta_2\pi^{-1}$. The ψ and h parameters are identified while the π parameter is identified if and only if $\beta_1 \neq 0$ and $\beta_2 \neq 0$. We assume the variance, σ^2, is positive, so that $\beta_1 = 0$ if and only if $\lambda_1 = 0$ and $\beta_2 = 0$ if and only if $\lambda_2 = 0$. This translates the rank condition on the factor loadings into the new parameters.

The objective function for this model is the likelihood implied by assuming joint normality on the errors and the factors. It is given by:

$$Q_n(\theta) = \log(|\Omega(\theta)|) + tr(S\Omega(\theta)^{-1}),$$

where $|\Omega|$ denotes the determinant of Ω, $tr(X)$ denotes the trace of a matrix, X, and S is the empirical covariance matrix,

$$S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'.$$

It is clear that $Q_n(\theta)$ depends on θ only through the reduced form parameters, $\delta = (\psi, h)$, because $\Omega(\theta)$ depends on θ only through the reduced form parameters. This shows that Example 1 satisfies the setup of Definition Parameter Spaces and Definition Objective Functions. □

In unidentified models, the shape of the parameter space becomes important. The following definitions are the relevant characteristics of the shape of the parameter space. In these definitions, the identified set and other objects may depend on the true value of the parameter, $\theta_0 = (\psi_0, \pi_0)$.

Definition 3.1.

(a) For any $\Psi \subset \mathbb{R}^{d_\psi}$, let $\Pi(\Psi) = \{\pi \in \mathbb{R}^{d_\pi} : (\psi, \pi) \in \Theta$ for some $\psi \in \Psi\}$.

(b) Let $\bar{\Pi} = \Pi(\mathbb{R}^{d_\psi})$.

(c) Let $\Pi_0(\psi_0) = \Pi(\{\psi_0\})$.

(d) For each $\pi \in \bar{\Pi}$, let $\Psi(\pi) = \{\psi \in \mathbb{R}^{d_\psi} : (\psi, \pi) \in \Theta\}$.

(e) For every $\eta > 0$, let $\Pi^{-\eta}(\psi_0) = \{\pi \in \Pi_0(\psi_0) : \inf_{\theta \in \partial\Theta} ||(\psi_0, \pi) - \theta|| \geq \eta\}$, where $\partial\Theta = cl(\Theta) \cap cl(\Theta^c)$.

Remarks:

1. Θ need not be defined as a product space between the ψ and π parameters. This means that cross sections of the parameter space, either in the ψ or the π direction may depend on the value of the other parameter. Part (a) gives notation for a π cross section that depends on a set of values of ψ.

2. Part (b) defines the largest cross section for π, which is all the values π can take.

3. Part (c) defines the identified set for π. When the value of ψ_0 is understood, this may be denoted Π_0. When π is not identified, its identified set stretches to the boundary of the parameter space. This can incorporate bounds on the parameter by including them in the definition of the parameter space.

4. Part (d) gives notation for the cross section in the ψ direction for a given value of π.

5. $\partial\Theta$, in part (e), denotes the boundary of the parameter space. Part (e) shrinks the identified set by a small amount, η, away from the boundary of the parameter space. This definition is helpful for dealing with nondifferentiability of the boundary of the identified set. When the true value of ψ_0 is understood, $\Pi^{-\eta}(\psi_0)$ is also denoted by $\Pi_0^{-\eta}$.

The following assumption provides regularity conditions on the shape of the parameter space.

Assumption HC.

(a) Θ *is compact.*

(b) $\bar{\Theta}$ *is relatively open in* $\mathbb{R}^{d_\psi} \times \bar{\Pi}$.

(c) $\bar{\Delta}$ *is open.*

(d) For any sequence of sets, Ψ_m*, converging in Hausdorff metric to* $\{\psi_0\}$, $\Pi(\Psi_m)$ *converges to* $\Pi_0(\psi_0)$ *in Hausdorff metric.*

(e) As $\eta \to 0$, $\Pi^{-\eta}(\psi_0)$ *converges in Hausdorff metric to* $\Pi_0(\psi_0)$.

Remarks:

1. HC stands for Hausdorff continuity, referring to part (d), which requires Hausdorff continuity of the boundary of the identified set for π.

2. Parts (a) and (b) together imply that there exist sets, $\Pi_0^+ \subset \mathbb{R}^{d_\pi}$ and $\Psi_0^+ \subset \mathbb{R}^{d_\psi}$, such that
$$\{\psi_0\} \times \Pi_0 \subset \text{int}\left(\Psi_0^+\right) \times \Pi_0^+ \subset \Psi_0^+ \times \Pi_0^+ \subset \bar{\Theta}.$$[19]

   Figure 2 illustrates the relationship between these sets. The figure plots the parameter space, Θ. The dark arrows are the boundary of Θ. Notice that the boundary, as a function of ψ, is allowed to depend on ψ and may not be differentiable. The dotted lines denote the boundary of $\bar{\Theta}$, the open expansion of Θ. The identified set, Π_0, are

[19] This statement is proved in Lemma 4.1(SM1) in Supplemental Materials 1.

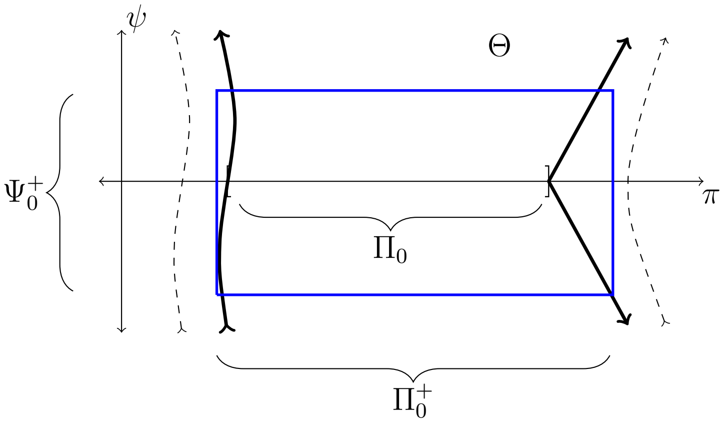

Figure 2: An illustration of the relationship between the identified set, Π_0, neighborhoods of the identified set, Π_0^+ and Ψ_0^+, and the parameter spaces, Θ and $\bar{\Theta}$.

the values of π that stretch from one boundary to the other when $\psi = \psi_0$, a point of identification failure. In the figure, $\psi_0 = 0$ is taken as an example. Finally, this figure illustrates the existence of sets Ψ_0^+ and Π_0^+ that are neighborhoods of the identified set and form a product space that is contained within $\bar{\Theta}$.

Defining Ψ_0^+ and Π_0^+ helps to deal with the boundary without assuming differentiability. $\Pi_0^+ \times \Psi_0^+$ is a neighborhood of the identified set that is extended globally for the unidentified parameters (π) and extended only locally for the identified parameters (ψ). This represents the fact that local properties of the objective function, that are sufficient for asymptotic analysis under identification, are insufficient for asymptotic analysis under weak or no identification. The global properties of the objective function that extend throughout the identified set must be taken into account. The theorems in this paper give the correct sense in which the global properties of the objective function contribute to the asymptotic theory.

3. Parts (b) and (c) give the sense in which the objective functions are defined on an enlargement of the parameter spaces.

4. Part (d) ensures continuity of the boundary of the parameter space in the π direction. It is important to note that this does not require differentiability of the boundary. This means that the boundary, and therefore the identified set for π, can be defined by the intersection of multiple bounds without violating this assumption.

5. Part (e) ensures the identified set is well behaved, and in particular does not shrink to a point. If the identified set were to shrink to a point, then the parameter would be identified due to the boundary. Asymptotics based on identification due to the

boundary are different from asymptotics based on identification due to the objective function. In order to focus on the latter, this assumption eliminates the influence of the boundary.

Example 1, Continued. When $\beta_1 = 0$ or $\beta_2 = 0$, π is still partially identified by bounds. These bounds come from the nonnegativity of the variance parameters. For example, $\phi_1 \geq 0$, the variance of the first error, gets reparameterized into $\pi \leq \zeta_1$. Since ζ_1 is identified, this defines an upper bound for π. Another example is $\phi_2 \geq 0$, the variance of the second error, gets reparameterized into $\pi \geq \frac{\beta_1^2}{\zeta_2}$.[20] The identified set for π is the interval defined by the minimum of all the upper bounds and the maximum of all the lower bounds. In Example 1, this is

$$\Pi_0 = \left[\max\left(\frac{\beta_1^2}{\zeta_2}, \frac{\beta_2^2}{\zeta_3}, \epsilon\right), \zeta_1\right].^{21}$$

The presence of the maximum means that the boundary of the identified set is not differentiable at some points. Example 1 illustrates the need to allow for the boundary of the identified set to depend on other parameters in a possibly nondifferentiable way in Assumption HC. □

Definition Parameter spaces, Definition Objective Functions, Definition 3.1, and Assumption HC define a class of models that are doubly parameterized by structural and reduced form parameters. A similar assumption, involving a double parameterization with structural and reduced form parameters, was used by Rothenberg (1971) to analyze identification in parametric models.

Showing that a given model fits into this structure typically requires two steps. The first is finding reduced form parameters that are always identified. Models that satisfy this first step include factor models, where the reduced form parameter is the covariance matrix of observables, curved exponential families, where the reduced form parameters are the coefficients in the exponent of the density, and minimum distance models, where the mapping from the structural to the reduced form parameters is given. More generally, arguments for identification in structural models tend to use this structure, stating that there are some reduced form parameters that are identified and these can be mapped back to the structural parameters, making this an intuitive and useful setup for dealing with identification failure in structural models. The second step is reparameterizing the structural parameters so that the ψ parameters determine the identification status of π. This can only be solved on a model by model basis and often requires some creativity. However, Han and McCloskey (2016) present a strategy for finding reparameterizations that may be helpful. Example 1

[20] The denominator, ζ_2 may be 0, but only when $\beta_1 = 0$, so the lower bound can always be well-defined.

[21] For technical reasons, π must be bounded away from 0 by some $\epsilon > 0$. The lower bounded is stated to reflect this. More details can be found in Supplemental Materials 2.

has already been reparameterized into this setup. Section 5 defines a second example and gives a reparameterization that satisfies this setup.

3.2 Theorem 1

This section provides a theorem for characterizing the asymptotic distribution of an extremum estimator within the class of models defined in Section 3.1. In particular, Theorem 1 holds for a sequence of parameters, $\theta_n = (\psi_n, \pi_n) \to \theta_0$, converging to a point where identification fails. For this section, fix one such sequence and all the probability statements (convergence in probability, convergence in distribution) hold under this sequence of true parameters.[22]

The extremum estimator is defined in two parts.

Definition EE1. For each $\pi \in \bar{\Pi}$, let

(a) $\hat{\psi}_n(\pi) \in \Psi(\pi)$ such that $Q_n(\hat{\psi}_n(\pi), \pi) \leq \inf_{\psi \in \Psi(\pi)} Q_n(\psi, \pi) + \tau_n$,

(b) $Q_n^c(\pi) = Q_n(\hat{\psi}_n(\pi), \pi)$,

(c) $\hat{\pi}_n \in \bar{\Pi}$ such that $Q_n^c(\hat{\pi}_n) \leq \inf_{\pi \in \bar{\Pi}} Q_n^c(\pi) + \tau_n$, and

(d) $\hat{\theta}_n = (\hat{\psi}_n(\hat{\pi}_n), \hat{\pi}_n)$,

where τ_n does not depend on π and satisfies $\tau_n = o_p(n^{-1})$.

Remarks:

1. τ_n may be further restricted in the following assumptions.

2. For notation, we can let $\hat{\psi}_n = \hat{\psi}_n(\hat{\pi}_n)$.

3. We notice that with this definition, $Q_n(\hat{\theta}_n) \leq \inf_{\theta \in \Theta} Q_n(\theta) + 2\tau_n$ so that $\hat{\theta}_n$ satisfies a typical extremum estimator condition.

Next, we impose regularity conditions on $T_n(\delta)$ and $h(\psi, \pi)$. These conditions are verified in the two examples in Sections 4 and 5 and Supplemental Materials 2 and 3.

Assumption RF-ID.

(a) There exists a nonstochastic real-valued function $T(\delta)$, continuous on $\bar{\Delta}$, such that

$$\sup_{\delta \in \bar{\Delta}} |T_n(\delta) - T(\delta)| \to_p 0.$$

[22] Also fix a sequence of (possibly infinite dimensional) nuisance parameters that index $Pr_{n,\delta}$.

(b) For every neighborhood, Δ_0, of $\delta_0 = (\psi_0, h(\psi_0, \pi_0))$,

$$\inf_{\delta \in \Delta/\Delta_0} T(\delta) - T(\delta_0) > 0.$$

Remarks:

1. RF-ID stands for "Reduced Form - Identified" because part (b) implies that the reduced form parameters are identified.

2. Part (a) gives the deterministic limit of the objective function, unstandardized. Part (b) ensures that the reduced form parameters, δ, are cleanly identified by $T_n(\delta)$ in the limit. These assumptions are standard assumptions for proving consistency of an extremum estimator in the literature.

Assumption RF-ID is sufficient for consistent estimation of the reduced form parameters, δ. However, the structural parameters are still not consistently estimable. The following lemma gives a type of consistency result while allowing for unidentified π.

Lemma CON. *Under Assumptions HC and RF-ID, the following hold.*

(a) There exists a sequence of sets Π_n that converges in Hausdorff metric to Π_0 such that $Pr(\hat{\pi}_n \in \Pi_n) \to 1$.

(b) $\sup_{\pi \in \Pi_n} ||\hat{\psi}_n(\pi) - \psi_0|| = o_p(1)$.

Remarks:

1. Part (a) is a type of concentration result because there exists a rate at which the distribution of $\hat{\pi}$ concentrates on Π_0, the identified set, determined by the rate at which Π_n converges to Π_0 in Hausdorff metric.

2. Π_n helps to solve nondifferentiability of the boundary. It allows us to replace the true, fixed boundary, which may depend on the value of ψ in a possibly nondifferentiable way, by a well behaved, but drifting boundary. Π_n is illustrated in Figure 3, which is a variant of Figure 2. As the sample size increases, Π_n converges to Π_0.

3. Part (b) is a uniform consistency result for $\hat{\psi}_n(\pi)$. This result is somewhat delicate because for fixed π, $\hat{\psi}_n(\pi)$ converges in probability to ψ_0 if and only if $\pi \in \Pi_0$. For this reason, the uniformity is taken over Π_n.

4. Parts (a) and (b) together show that $(\hat{\psi}_n(\pi), \hat{\pi}_n) \in \Psi_0^+ \times \Pi_0^+$ for all $\pi \in \Pi_n$ with probability approaching 1.

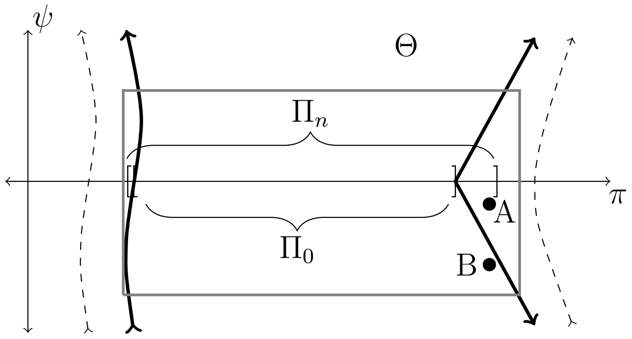

Figure 3: An illustration of the identified set, Π_0, a drifting expansion of the identified set, Π_n, and the role of the boundary of Θ when optimizing over ψ for fixed $\pi \notin \Pi_0$.

The hat estimators can now be thought of as extremum estimators over $\Psi_0^+ \times \Pi_0^+$ constrained by the (possibly nondifferentiable) boundary of Θ. We can also define the unconstrained estimators.

Definition EE2. For each $\pi \in \Pi_0^+$, let

(a) $\tilde{\psi}_n(\pi) \in \Psi_0^+$ such that $Q_n(\tilde{\psi}_n(\pi), \pi) \leq \inf\limits_{\psi \in \Psi_0^+} Q_n(\psi, \pi) + \tau_n$,

(b) $\tilde{Q}_n^c(\pi) = Q_n(\tilde{\psi}_n(\pi), \pi)$, and

(c) $\tilde{\pi}_n \in \Pi_n$ such that $\tilde{Q}_n^c(\tilde{\pi}_n) \leq \inf\limits_{\pi \in \Pi_n} \tilde{Q}_n^c(\pi) + \tau_n$.

Remark:

1. Figure 3 illustrates why $\tilde{\psi}_n(\pi)$ can be considered an unconstrained estimator. $\tilde{\psi}_n(\pi)$ is allowed to take any value within Ψ_0^+ and is not restricted by the boundary of Θ when $\pi \notin \Pi_0$. For example, consider the value of π on the right of Figure 3 that is common to points A and B. $\tilde{\psi}_n(\pi)$ is allowed to optimize over the whole range of Ψ_0^+ and may find a minimum at point A. In contrast, $\hat{\psi}_n(\pi)$ can only optimize over points in Θ, and is forced to find a higher minimum, possibly at point B.

Lemma CON. *Under Assumptions HC and RF-ID,*

(c) $\sup\limits_{\pi \in \Pi_0^+} ||\tilde{\psi}_n(\pi) - \psi_0|| = o_p(1)$.

Remark:

1. Without the constraint imposed by the boundary, we can prove part (c), a stronger uniform consistency result that holds over all of Π_0^+.

Now that we have consistency, we impose assumptions on $T_n(\delta)$ and $h(\psi, \pi)$ to characterize the asymptotic distribution.

Assumption h1.

(a) $h(\psi, \pi)$ is continuously differentiable in ψ, uniformly in $\pi \in \Pi_0^+$. Let $h^1(\psi, \pi)$ denote the $d_h \times d_\psi$ matrix of partial derivatives of $h(\psi, \pi)$ with respect to ψ. Also let $d_1(\pi) = h^1(\psi_0, \pi)$.

(b) There exists a $\mathbb{R}^{d_h}$-valued function, $d_0(\pi)$, on Π_0^+ such that $\sqrt{n}(h(\psi_n, \pi) - h(\psi_n, \pi_n)) \to d_0(\pi)$ uniformly over $\pi \in \Pi_0^+$.

(c) $d_0(\pi)$ and $d_1(\pi)$ are continuous in π.

Remarks on Assumption h1:

1. Part (a) means that $h^1(\psi, \pi)$ is continuous uniformly over π. That is, for every $\bar{\psi}_n \to \psi_0$, $h^1(\bar{\psi}_n, \pi) \to h^1(\psi_0, \pi)$ uniformly over $\pi \in \Pi_0^+$.
2. Part (b) give the defining characteristics of a weakly identified sequence of parameters—that $h(\psi_n, \pi)$ converges to zero at the $n^{-1/2}$ rate.

Assumption QE. *Let $\delta_n = \delta(\theta_n)$. For every $\delta = (\psi, h) \in \bar{\Delta}$, the following holds.*

(a) The reduced form objective function $T_n(\psi, h)$ satisfies a quadratic expansion in $\delta = (\psi, h)$ around δ_n:

$$T_n(\delta) \quad = \quad T_n(\delta_n) + D^1 T_n(\delta_n)'(\delta - \delta_n) + \frac{1}{2}(\delta - \delta_n)' D^2 T_n(\delta_n)(\delta - \delta_n) + R_2^T(\delta),$$

where $D^1 T_n(\delta)$ denotes a generalized first derivative vector and $D^2 T_n(\delta)$ denotes a generalized second derivative matrix that may be stochastic or nonstochastic.

(b) The remainder, $R_2^T(\delta)$, satisfies, for all constants $\eta_n \to 0$,

$$\sup_{\delta \in \bar{\Delta}: ||\delta - \delta_n|| \leq \eta_n} \frac{|n R_2^T(\delta)|}{(1 + ||\sqrt{n}(\delta - \delta_n)||)^2} = o_p(1).$$

Assumption T1. *There exists a random vector, Y_1, and a $d_\delta \times d_\delta$ matrix, $\bar{V}$, such that*

$$\sqrt{n} D^1 T_n(\delta_n) \to_d Y_1 = \begin{pmatrix} Y_{1\psi} \\ Y_{1h} \end{pmatrix} \sim N(0, \bar{V}).$$

Assumption T2. *There exists an $H \in \mathbb{R}^{d_\delta \times d_\delta}$, positive definite and symmetric, such that $D^2 T_n(\delta_n) \to_p H$.*

Remarks on Assumptions QE, T1, and T2:

1. QE stands for quadratic expansion. T1 stands for the limit of the first derivative of $T_n(\delta)$. T2 stands for the limit of the second derivative of $T_n(\delta)$.

2. Assumption QE is a standard quadratic expansion around the true value. Sufficient conditions for Assumption QE are that $Q_n(\theta)$ is twice continuously differentiable almost surely and the second derivative is stochastically equicontinuous.

3. Assumption T1 can be verified by a central limit theorem. We allow $\bar{V} = \bar{V}(\delta_0)$ to depend on the true limit of the sequence.

4. Assumption T2 ensures that $D^2 T_n(\delta_n)$ is eventually positive definite.

5. Assumption h1 and T1 imply that the first derivative of the structural objective function is equal to

$$D^1 Q_n(\theta_n) = \begin{bmatrix} I_{d_\psi} & 0_{d_\psi \times d_\pi} \\ h^1(\psi_n, \pi_n) & \dfrac{\partial}{\partial \pi'} h(\psi_n, \pi_n) \end{bmatrix}' D^1 T_n(\delta(\theta_n)).$$

   When $h(\psi, \pi)$ is not invertible for π, $\dfrac{\partial}{\partial \pi'} h(\psi, \pi)$ is zero. This implies that for weak sequences, the covariance matrix of $D^1 Q_n(\theta_n)$ is singular in the limit. This is the first order mechanism through which weak identification affects asymptotics.

Example 1, Continued. Assumptions QE, T1, and T2 are easy to verify in Example 1. They rely on the smoothness of the objective function and the asymptotic normality of S, the empirical covariance matrix.

Assumption h1 holds for some sequences, θ_n, in Example 1. For $h(\psi, \pi) = \beta_1 \beta_2 \pi^{-1}$, the important property of a sequence is the rate at which β_{1n} and β_{2n} converge to zero. Any sequence such that $\beta_{1n}\beta_{2n}$ converges to zero at an $n^{-1/2}$ rate or faster satisfies Assumption h1. If $\sqrt{n}\beta_{1n}\beta_{2n} \to b$ for some $b \in \mathbb{R}$, then

$$\begin{aligned} d_0(\pi) &= b(\pi^{-1} - \pi_0^{-1}), \text{ and} \\ d_1(\pi) &= \begin{pmatrix} \beta_{20}\pi^{-1} & \beta_{10}\pi^{-1} & 0 & 0 & 0 \end{pmatrix}, \end{aligned}$$

where $(\beta_{1n}, \beta_{2n}) \to (\beta_{10}\beta_{20})$, at least one of which is zero. This shows that Assumptions h1, QE, T1, and T2 are easy to verify in Example 1. □

Partition $D^1 T_n(\delta_n)$ into $(D_\psi T_n(\delta_n)', D_h T_n(\delta_n)')'$. Also partition

$$D^2 T_n(\delta_n) = \begin{bmatrix} D_{\psi\psi} T_n(\delta_n) & D_{\psi h} T_n(\delta_n) \\ D_{h\psi} T_n(\delta_n) & D_{hh} T_n(\delta_n) \end{bmatrix}$$

and

$$H = \begin{bmatrix} H_{\psi\psi} & H_{\psi h} \\ H_{h\psi} & H_{hh} \end{bmatrix}.$$

We characterize the distribution of $\hat{\psi}_n(\pi)$ and $\tilde{\psi}_n(\pi)$ as a function of π. Let

$$Z(\pi) = -\left(\begin{bmatrix} I_{d_\psi} \\ d_1(\pi) \end{bmatrix}' H \begin{bmatrix} I_{d_\psi} \\ d_1(\pi) \end{bmatrix} \right)^{-1} \begin{bmatrix} I_{d_\psi} \\ d_1(\pi) \end{bmatrix}' \left(Y_1 + \begin{bmatrix} H_{\psi h} \\ H_{hh} \end{bmatrix} d_0(\pi) \right)$$

be a stochastic processes defined on Π_0^+.

Lemma $\boldsymbol{\psi}$ Limit. *Under Assumptions HC, RF-ID, h1, QE, T1, T2, and for all $\eta > 0$,*

(a) $\sqrt{n}(\tilde{\psi}_n(\pi) - \psi_n) \Rightarrow Z(\pi)$ as a stochastic process over Π_0^+, and

(b) $\sqrt{n}(\hat{\psi}_n(\pi) - \psi_n) \Rightarrow Z(\pi)$ as a stochastic process over $\Pi_0^{-\eta}$.

Remarks:

1. This is called Lemma ψ Limit because it characterizes the limiting distribution for the estimators of ψ.

2. $\tilde{\psi}_n(\pi)$ has a Gaussian stochastic process limit for all values of $\pi \in \Pi_0^+$. In contrast, $\hat{\psi}_n(\pi)$ does not have a Gaussian limit for $\pi \notin \text{int}(\Pi_0)$, so the Gaussian stochastic process must be truncated short of the boundary. This is why the convergence in (b) holds only over $\Pi_0^{-\eta}$ instead of the full set, Π_0^+.

The following definition gives the limit for the concentrated objective function. Let

$$\xi(\pi) = Y_{1h}' d_0(\pi) + \frac{1}{2} d_0(\pi)' H_{hh} d_0(\pi) - \frac{1}{2} Z(\pi)' \left(\begin{bmatrix} I_{d_\psi} \\ d_1(\pi) \end{bmatrix}' H \begin{bmatrix} I_{d_\psi} \\ d_1(\pi) \end{bmatrix} \right) Z(\pi) \tag{3.1}$$

be a stochastic process defined on Π_0^+.

Lemma $\boldsymbol{Q}$ Limit. *Under Assumptions HC, RF-ID, h1, QE, T1, and T2,*

(a) $n(\tilde{Q}_n^c(\pi) - T_n(\delta_n)) \Rightarrow \xi(\pi)$ as a stochastic process on Π_0^+, and

(b) for every $\eta > 0$. $n(Q_n^c(\pi) - T_n(\delta_n)) \Rightarrow \xi(\pi)$ at a stochastic process on $\Pi_0^{-\eta}$.

Assumption MIN. *Almost surely, each sample path of the stochastic process $\xi(\pi)$ is continuous over Π_0 and is uniquely minimized over Π_0 at a unique point denoted π_{MIN}.*

Remarks:

1. Assumption MIN is a high level condition, but equation (3.1) gives a formula for $\xi(\pi)$ in terms of primitives. In addition, Cox (2016b) provides sufficient conditions for verifying Assumption MIN based on a type of transversality condition on $\xi(\pi)$.

2. It is clear from the formula for $\xi(\pi)$ in equation (3.1) that Assumption MIN is not satisfied whenever both $d_0(\pi)$ and $d_1(\pi)$ do not depend on π. For these cases, Theorem 2, in Section 3.3, allows a much weaker version of Assumption MIN.

3. Lemma Q Limit(a) together with Assumption MIN are sufficient to characterize the asymptotic distribution of $\tilde{\pi}_n$ using the argmax theorem.[23] However, Lemma Q Limit(b) is insufficient to characterize the asymptotic distribution of $\hat{\pi}_n$, because the convergence holds only over the smaller set, $\Pi_0^{-\eta}$.

Example 1, Continued. In Example 1, there exist weak sequences, θ_n, such that $\sqrt{n}\beta_{1n}\beta_{2n} \to 0$ and $\beta_{10} = \beta_{20} = 0$. For these sequences, $d_0(\pi) = 0$ and $d_1(\pi) = 0_{1\times 5}$. This implies that $\xi(\pi)$ does not depend on π and, in particular, does not satisfy Assumption MIN. These sequences are called "super-weak" and cannot be handled by Theorem 1. Instead, they are handled by Theorem 2 in Section 4.3. However, for most weak sequences, $\sqrt{n}\beta_{1n}\beta_{2n} \to b \neq 0$ or $(\beta_{1n}, \beta_{2n}) \to (\beta_{10}, \beta_{20}) \neq (0,0)$, showing that $\xi(\pi)$ is nondegenerate, and these sequences can be handled by Theorem 1. □

The following lemma allows us to connect the constrained estimator, $\hat{\psi}_n(\pi)$, to the unconstrained estimator, $\tilde{\psi}_n(\pi)$, when evaluated at $\hat{\pi}_n$.

Lemma B. *Under Assumptions HC, RF-ID, h1, QE, T1, and T2,*

(a) $\hat{\psi}_n(\hat{\pi}_n) = \tilde{\psi}_n(\hat{\pi}_n) + o_p(n^{-1/2})$.

(b) If, in addition, Assumption MIN holds, $\hat{\pi}_n = \tilde{\pi}_n + o_p(1)$.

Remarks:

1. B stands for boundary because this lemma shows that the constraint imposed by the boundary in the definition of $\hat{\psi}_n(\pi)$ is negligible for the limit theory.

2. This result is surprising. Lemma ψ Limit shows that $\hat{\psi}_n(\pi)$ and $\tilde{\psi}_n(\pi)$ are close when on the interior of Π_0. However, this is not enough because there is a positive probability in the limit that $\hat{\pi}_n$ is not in the interior of Π_0. When $\pi \notin \text{int}(\Pi_0)$, $\hat{\psi}_n(\pi)$ and $\tilde{\psi}_n(\pi)$ have very different behavior (because $\hat{\psi}_n(\pi)$ does not converge). Lemma B shows that this different behavior is irrelevant when evaluated at $\hat{\pi}_n$.

[23] See van der Vaart and Wellner (1996), Section 3.2.1.

Theorem 1. *Under Assumptions HC, RF-ID, h1, QE, T1, T2, and MIN,*

$$\left(\begin{array}{c} n^{1/2}(\hat{\psi}_n(\hat{\pi}_n) - \psi_n) \\ \hat{\pi}_n \end{array} \right) \rightarrow_d \left(\begin{array}{c} Z(\pi_{MIN}) \\ \pi_{MIN} \end{array} \right).$$

Remarks:

1. Theorem 1 provides the asymptotic limit theory for $\hat{\theta}_n$. The asymptotic distribution of $\hat{\psi}$ can be characterized as a Gaussian stochastic process function of π_{MIN}. The asymptotic distribution of $\hat{\pi}$ can be characterized as the argmin of a stochastic process, π_{MIN}. Notice that the rate for $\hat{\psi}_n$ is $n^{1/2}$ while $\hat{\pi}_n$ is inconsistent and converges unstandardized.

2. Theorem 1 relies on Assumption MIN, which assumes n is the correct rate to standardize $Q_n^c(\pi)$. For some sequences, $Q_n^c(\pi)$ may need to be standardized at a faster rate. Theorem 2 in the next section handles those cases.

3.3 Theorem 2

Section 3.2 states a theorem for characterizing the asymptotic distribution of an extremum estimator that relies on Assumption MIN. This section provides a theorem that relies on a weaker version of that assumption, Assumption MIN*.

The key is to recognize that the objective function can be restandardized. At faster rates, higher-order terms become relevant for the limit theory. The following assumptions make these higher-order terms explicit.

The following assumption imposes conditions on higher-order derivatives of $h(\psi, \pi)$ with respect to ψ. The notation for this requires higher dimensional matrices or tensors. Let $h^m(\psi, \pi)$ denote the $d_h \times d_\psi \times ... \times d_\psi$ tensor of partial derivatives of $h(\psi, \pi)$ with respect to ψ.[24]

Assumption h2. *There exists a $K > 1$ such that the following hold.*

(a) $h(\psi, \pi)$ is K-times continuously differentiable with respect to ψ, uniformly over π. Let $d_K^*(\pi) = h^K(\psi_0, \pi)$.

(b) There exists a sequence of positive constants, a_n^0, and an $\mathbb{R}^{d_h}$-valued function on Π_0^+, $d_0^*(\pi)$, such that $a_n^0(h(\psi_n, \pi) - h(\psi_n, \pi_n)) \to d_0^*(\pi)$ uniformly in π and $n^{-1/2}a_n^0 \to \infty$.

(c) For each $m = 1, ..., K-1$, there exists a sequence of positive constants, $a_n^m \to \infty$, and a $d_h \times d_\psi \times ... \times d_\psi$-tensor valued function on Π_0^+, $d_m^*(\pi)$, such that $a_n^m h^m(\psi_n, \pi) \to d_m^*(\pi)$, uniformly in π.

[24] A description of the notation for dealing with tensors can be found in Appendix D.

(d) $d_0^*(\pi)$ and $d_m^*(\pi)$ are continuous in π for each $m = 1, ..., K$.

Remarks on Assumption h2:

1. The uniform continuous differentiability in part (a) implies that for any $\bar{\psi}_n \to \psi_0$, $\sup_{\pi \in \Pi_0^+} ||h^K(\bar{\psi}_n, \pi) - d_K^*(\pi)|| \to 0$. We can let $a_n^K = 1$ for all n.

2. Parts (b) and (c) give the defining characteristic of a super-weak sequences of parameters—that $h(\psi_n, \pi)$ converges to zero faster than $n^{-1/2}$ and the first derivative of $h(\psi_n, \pi)$ also converges to zero.

The next assumption imposes a higher degree of smoothness on the reduced form objective function, allowing an expansion out to a higher order.

Assumption T3+.

(a) The reduced form objective function $T_n(\delta)$ is $K+1$ times continuously differentiable in δ almost surely. Denote the k^{th} partial derivative matrix by $D^k T_n(\delta)$, which is $d_\delta \times d_\delta \times \cdots \times d_\delta$ with k dimensions, for $k = 1, ..., K+1$.

(b) For each $k = 3, ..., K+1$, there exists a sequence of positive constants, $\tilde{a}_n^k$, such that $\liminf_{n\to\infty} \tilde{a}_n^k > 0$ and

$$\tilde{a}_n^k D^k T_n(\delta_n) \to_d Y_k.$$

(c) $D^{K+1} T_n(\delta)$ is stochastically equicontinuous at δ_0. That is, for every $\epsilon > 0$ there exists an $\iota > 0$ such that

$$\limsup_{n\to\infty} Pr\left(\sup_{\delta : ||\delta - \delta_0|| \le \iota} ||D^{K+1} T_n(\delta) - D^{K+1} T_n(\delta_0)|| > \epsilon\right) < \epsilon.$$

Remarks:

1. Also note that we can define $\tilde{a}_n^1 = \sqrt{n}$ and $\tilde{a}_n^2 = 1$ because of Assumptions T1 and T2.

2. Note that we only need to assume that $D^{K+1} T_n(\delta)$ is stochastically equicontinuous and not $\tilde{a}_n^{K+1} D^{K+1} T_n(\delta_n)$, which would be stronger.

3. The norm in part (c) is the Euclidean norm for tensors (see Appendix D).

Example 1, Continued. In Example 1, there exist sequences, θ_n, such that $\sqrt{n}\beta_{1n}\beta_{2n} \to 0$ and $(\beta_{1n}, \beta_{2n}) \to (0, 0)$. These are sequences that could not be handled by Theorem 1. For these sequences, Assumption h2 is satisfied by taking $K = 2$. Then, part (a) is satisfied by a simple calculation that shows that $h^2(\psi, \pi)$ is a nonzero ($1 \times 5 \times 5$-tensor valued) multiple of π^{-1}. Part (b) holds because $h(\psi_n, \pi)$ converges to zero at the $\beta_{1n}\beta_{2n}$ rate, which is faster than $\sqrt{n}$. Part (c) holds because $h^1(\psi_n, \pi)$ converges to zero at the $\max(\beta_{1n}, \beta_{2n})$ rate.

The advantage that Assumption h2 has over Assumption h1 in Example 1 is the fact that some $d^*_m(\pi)$ is guaranteed to depend on π, no matter how quickly β_{1n} and β_{2n} converge to zero. In particular, $d^*_2(\pi)$ is a nonzero multiple of π^{-1} which can be used to provide an upper bound on the rate of convergence of $Q^c_n(\pi)$ to a nondegenerate limit.

Assumption T3+ is easily satisfied in Example 1. For $k = 3$, the third derivative of the reduced form objective function converges for $\tilde{a}^3_n = 1$. An expression for the limit can be found in Supplemental Materials 2. □

Assumption T3+ implies a polynomial expansion on the reduced form objective function.

Lemma PE. *Under Assumptions T1, T2, and T3+, the following hold.*

(a) The reduced form objective function $T_n(\psi, h)$ has a $K + 1$ order polynomial expansion in $\delta = (\psi, h)$ around δ_n:

$$T_n(\delta) \quad = \quad T_n(\delta_n) + \sum_{k=1}^{K+1} \frac{1}{k!} \langle D^k T_n(\delta_n); (\delta - \delta_n)^{\otimes k} \rangle + R^T_{K+1}(\delta).$$

(b) The remainder $R^T_{K+1}(\delta)$ satisfies, for all constants $\eta_n \to 0$,

$$\sup_{\delta \in \bar{\Delta}: ||\delta - \delta_n|| \leq \eta_n} \frac{|n^{(K+1)/2} R^T_{K+1}(\delta)|}{(1 + ||\sqrt{n}(\delta - \delta_n)||)^{K+1}} = o_p(1).$$

Remarks:

1. PE stands for polynomial expansion. This lemma is a generalization of Assumption QE from a quadratic expansion to a polynomial expansion of an arbitrary order.

2. The $\langle \cdot ; \cdot \rangle$ and $\otimes$ that appears in condition (a) are a tensor notation. The $\langle \cdot ; \cdot \rangle$ denotes an inner product on tensors and the $\otimes$ denotes a tensor product. More details on this notation can be found in Appendix D.

3. Lemma PE is similar to other higher-order expansions that have been used in the literature when there is degeneracy in the first order limit.[25]

In addition to the expansion of the objective function, I consider an expansion of the first order conditions (FOC). The next assumption ensures that the FOC hold at a fast enough rate.

[25] For other examples, see Sargan (1983) and Cho and White (2007).

Assumption FOC. *For every $\eta > 0$,*

$$\begin{aligned}\sup_{\pi \in \Pi_0^+} \|\frac{\partial}{\partial \psi} Q_n(\tilde{\psi}_n(\pi), \pi)\| &= o_p(n^{-K/2}), \text{ and} \\ \sup_{\pi \in \Pi_0^{-\eta}} \|\frac{\partial}{\partial \psi} Q_n(\hat{\psi}_n(\pi), \pi)\| &= o_p(n^{-K/2}).\end{aligned}$$

Also assume that $\tau_n = o_p(n^{-(K+1)/2})$.

Remark:

1. Assumption FOC can be satisfied either by setting $\tau_n = 0$ and arguing that FOC holds exactly, or by defining the extremum estimator to satisfy the FOC at the given rate, a common criterion for numerical optimization.

The above assumptions allow an expansion of the concentrated objective function into many terms, some of which depend on π and some do not. The next assumption provides a sufficient condition that among all these terms, there exists one term or a collection of terms that depends on π at the slowest rate.

Assumption Rates. *For every collection of constants $i_0, i_1, ..., i_{K-1} \in \mathbb{Z}$ and for every $j_1, ..., j_{K+1} \in \mathbb{Z}$, the sequence of positive constants defined by*

$$\breve{a}_n = (a_n^0)^{i_0}(a_n^1)^{i_1} \cdots (a_n^{K-1})^{i_{K-1}}(\tilde{a}_n^1)^{j_1} \cdots (\tilde{a}_n^{K+1})^{j_{K+1}}$$

satisfies $\limsup_{n\to\infty} \breve{a}_n = \liminf_{n\to\infty} \breve{a}_n$.

Remarks:

1. This assumption is much stronger than necessary. There are only a couple of places in the proof where this assumption is invoked, and each time there is only a finite number of potential sequences. Thus, a subsequencing argument can ensure that the necessary limits exist, so that this assumption is always satisfied.

2. Assumption Rates is satisfied whenever a_n^i and $\tilde{a}_n^j$ are powers of n, which occurs in many, but not all, cases.

The above assumptions are sufficient to characterize a limit for the the concentrated objective function. The faster rate of convergence means that this limit depends on π even when $\xi(\pi)$ from Lemma Q Limit does not.

Lemma Q Limit*. *Under Assumptions HC, RF-ID, h2, T1, T2, T3+, FOC, and Rates, there exists a sequence of random variables, $Q_{0,n}$, a sequence of positive constants $a_n^* \to \infty$, and a stochastic process over Π_0^+, $\xi^*(\pi)$, such that* *(a)* $a_n^*(\tilde{Q}_n^c(\pi) - Q_{0,n}) \Rightarrow \xi^*(\pi)$ *as a sequence of stochastic processes on* Π_0^+*, and*

(b) for every $\eta > 0$*,* $a_n^*(Q_n^c(\pi) - Q_{0,n}) \Rightarrow \xi^*(\pi)$ *as a sequence of stochastic processes on* $\Pi_0^{-\eta}$*.*

Assumption MIN*. *Almost surely, each sample path of the stochastic process* $\xi^*(\pi)$ *is continuous over* Π_0 *and is uniquely minimized over* Π_0 *at a unique point denoted* π^*_{MIN}*.*

Remarks on Lemma Q Limit* and Assumption MIN*:

1. Lemma Q Limit* and Assumption MIN* must be satisfied together. The purpose of this section is to bring higher-order terms into the limit in Lemma Q Limit* so that they can help to satisfy Assumption MIN*.

2. For a given sequence, the proof of Lemma Q Limit* can be followed to give a closed form solution for $\xi^*(\pi)$. Then, Assumption MIN* can be verified (1) directly, by finding a solution for π^*_{MIN}, or (2) indirectly, by proving the existence of a unique minimum almost surely. Cox (2016b) provides sufficient conditions for verifying the unique minimum condition indirectly.

Theorem 2. *Under Assumptions HC, RF-ID, h2, T1, T2, T3+, FOC, Rates, MIN*, and assuming that* $a_n^*\tau_n = o_p(1)$*,*

$$\begin{pmatrix} n^{1/2}(\hat{\psi}_n(\hat{\pi}_n) - \psi_n) \\ \hat{\pi}_n \end{pmatrix} \rightarrow_d \begin{pmatrix} Z(\pi^*_{MIN}) \\ \pi^*_{MIN} \end{pmatrix}.$$

Remarks:

1. The condition $a_n^*\tau_n = o_p(1)$ ensures that the original tolerance for minimization was sufficiently small.

2. The result for Theorem 2 is the same as that for Theorem 1, except with π_{MIN} replaced by π^*_{MIN}. This represents the fact that the standard limit for the objective function, $\xi(\pi)$ is degenerate and does not satisfy a unique min condition. Instead, the proof of this theorem uses a nonstandard limit, $\xi^*(\pi)$, that involves higher-order terms that are nondegenerate and can be used to define a limiting distribution for $\hat{\pi}_n$.

4 Robust Inference in Example 1

This section provides a method for robust inference in Example 1. Consider testing the null hypothesis $H_0 : r(\theta) = \nu$ using the Wald test statistic, $W_n(\nu)$.[26] This section proposes

[26] Likelihood ratio and Lagrange multiplier test statistics can also be used for some hypotheses, see Supplemental Materials 2 and 3 for examples.

testing H_0 robustly using a two step procedure. In the first step, the rank condition is tested to determine whether or not it is close to failing. If the rank condition is close to failing, $W_n(\nu)$ is compared to a robust critical value. If the rank condition is not close to failing, $W_n(\nu)$ is compared to a standard chi-squared critical value. The robust critical value is calculated based on the asymptotic distributions of $W_n(\nu)$ along a classes of sequences of parameters.

4.1 Asymptotic Distributions

This section divides sequences of parameters converging to points of rank condition failure into classes and characterizes the asymptotic distribution of the estimator and test statistic along sequences in each class.

The classes are defined based on the rate at which $h(\psi_n, \pi) = \beta_{1n}\beta_{2n}\pi^{-1}$ and $h_1(\psi_n, \pi) = (\beta_{2n}\pi^{-1}, \beta_{1n}\pi^{-1}, 0, 0, 0)$ converge. This only depends on the rates at which β_{1n} and β_{2n} converge to zero.

These classes are illustrated using a heuristic sketch, depicted in Figure 4. A point in the Northeast corner, for example point A, depicts a sequence, θ_n, that converges to a limit, θ_0, such that both $\beta_{10} > 0$ and $\beta_{20} > 0$, and analogously for the other corners. A point in the East area, for example point B, depicts a sequence, θ_n, that converges to a limit, θ_0, for which $\beta_{10} > 0$ and $\beta_{20} = 0$. Points that are further from the axis depict sequences for which β_{2n} converges slower to $\beta_{20} = 0$. A point in the central area, for example point C, depicts a sequence, θ_n, that converges to a limit, θ_0, for which $\beta_{10} = 0$ and $\beta_{20} = 0$. The relative position within the square determines the relative rates at which β_{1n} and β_{2n} converge to their limits, $\beta_{10} = 0$ and $\beta_{20} = 0$.

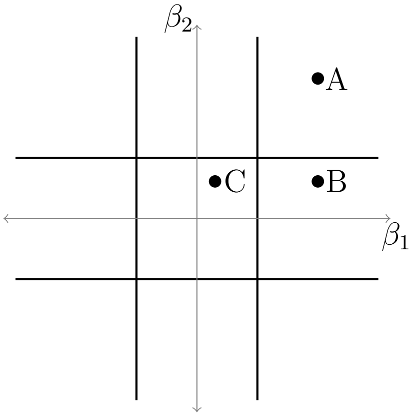

Figure 4

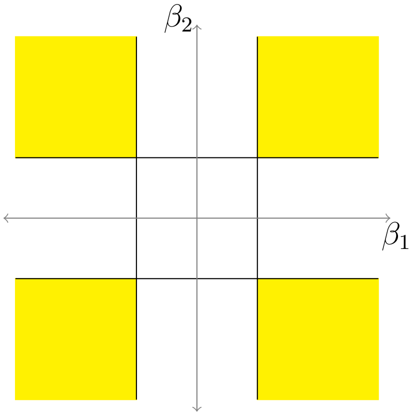

Figure 5

First, we can define the class of strong sequences, θ_n, as those sequences that converge to a limit, θ_0, such that both $\beta_{10} \neq 0$ and $\beta_{20} \neq 0$. These sequences are depicted in Figure 5 in yellow. Theorem 4, in Supplemental Materials 1, applied to strong sequences, gives an asymptotic normal distribution for $\hat{\theta}_n$.[27] The continuous mapping theorem implies that $W_n(\nu)$ converges to a limiting $\chi^2_{d_r}$ distribution.

[27] Supplemental Materials 2 verifies the assumptions of Theorem 4. More generally, Supplemental Materials 2 contains details and proofs for the assertions in this section (Section 4).

The class of semi-strong sequences is characterized by $h(\psi_n, \pi)$ converging to zero at a rate slower than $n^{-1/2}$. This occurs when

1. $\beta_{10} \neq 0$, $\beta_{20} = 0$, and $\sqrt{n}|\beta_{2n}| \to \infty$,
2. $\beta_{10} = 0$, $\beta_{20} \neq 0$, and $\sqrt{n}|\beta_{1n}| \to \infty$, or
3. $\beta_{10} = 0$, $\beta_{20} = 0$, and $\sqrt{n}|\beta_{1n}\beta_{2n}| \to \infty$.

These sequences are depicted in Figure 6 in green. The first category is depicted in the East and West areas, the second category is depicted in the North and South areas, and the third category is depicted by the curved part in the center. Theorem 3, in Supplemental Materials 1, applied to semi-strong sequences, gives

$$\begin{pmatrix} \sqrt{n}(\hat{\psi}_n - \psi_n) \\ \sqrt{n}a_n^{-1}(\hat{\pi}_n - \pi_n) \end{pmatrix} \to_d Z_S,$$

where $a_n \to \infty$ and Z_S is a normal random vector. a_n diminishes the rate of convergence of $\hat{\pi}_n$. The continuous mapping theorem implies that $W_n(\nu)$ converges to a limiting $\chi^2_{d_r}$ distribution.

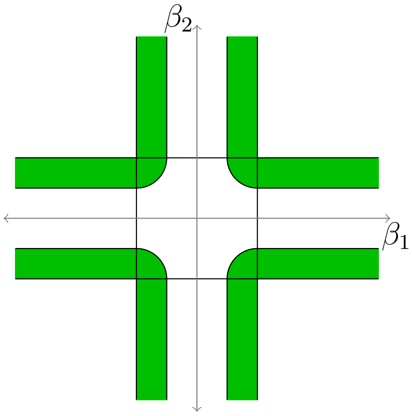

Figure 6

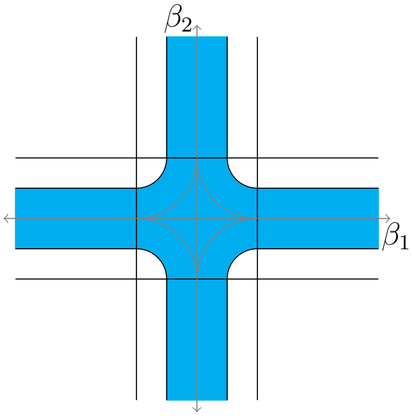

Figure 7

Weak sequences are characterized by $h(\psi_n, \pi)$ converging to zero at an $n^{-1/2}$ rate, or faster. For Example 1, there are three classes of weak sequences, defined by:

NS: $\beta_{10} = 0$, $\beta_{20} \neq 0$, and $\sqrt{n}\beta_{1n} \to b_1$,

EW: $\beta_{10} \neq 0$, $\beta_{20} = 0$, and $\sqrt{n}\beta_{2n} \to b_2$, or

C: $\beta_{10} = 0$, $\beta_{20} = 0$, and $\sqrt{n}\beta_{1n}\beta_{2n} \to b$,

where b_1, b_2, and b are local parameters indexing the rate of convergence. These sequences are depicted in Figure 7 in light blue. The NS class is depicted in the North and South areas of the picture, the EW class is depicted in the East and West areas of the picture, and the C class is depicted in the center of the picture. If we assume that $b \neq 0$ (b_1 and b_2 are allowed to take any real number, including 0), we can apply Theorem 1 to these sequences, which gives

$$\begin{pmatrix} \sqrt{n}(\hat{\psi} - \psi_n) \\ \hat{\pi} \end{pmatrix} \to_d \begin{pmatrix} Z_j(\pi_j^*) \\ \pi_j^* \end{pmatrix},$$

where $Z_j(\cdot)$ is a Gaussian stochastic process over π, π_j^* is a random variable whose distribution is characterized as the argmin of $\xi(\pi)$, and $j \in \{NS, EW, C\}$.[28] This asymptotic distribution is indexed by (1) the value of the local parameter, b, b_1, or b_2, and (2) the value of the true parameter, θ_0. We can group the local parameters and θ_0 into a single index, called $g \in G_j$, where G_j is a collection of indices. The continuous mapping theorem implies that $W_n(\nu)$ converges to a nonstandard limit distribution, $F_j(g)$, for $g \in G_j$.

When $b = 0$, neither $d_0(\pi)$ nor $d_1(\pi)$ depends on π. This implies that Assumption MIN is not satisfied for these sequences because the limit of the concentrated objective function, $\xi(\pi)$, is degenerate in the sense that it does not depend on π. The solution is to use Theorem 2, restandardizing the objective function and keeping track of higher-order terms that do depend on π.

Super-weak sequences are weak sequences such that both 1. $h(\psi_n, \pi)$ converges to zero at a rate faster than $n^{-1/2}$ and 2. $h_1(\psi_n, \pi)$ converges to zero. In Example 1, there are four classes of super-weak sequences. In addition to satisfying part C of the weak sequences definition for $b = 0$, they are defined by:

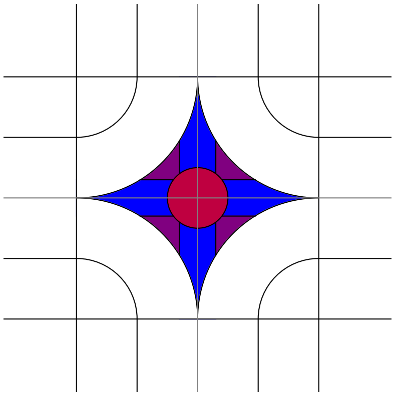

Figure 8

E: $\sqrt{n}|\beta_{1n}| \to \infty$ and $\sqrt{n}|\beta_{2n}| \to \infty$,

U1: $\sqrt{n}|\beta_{1n}| \to \infty$ and $\sqrt{n}\beta_{2n} \to b_2^*$,

U2: $\sqrt{n}\beta_{1n} \to b_1^*$ and $\sqrt{n}|\beta_{2n}| \to \infty$, or

CC: $\sqrt{n}(\beta_{1n}, \beta_{2n}) \to (b_1^*, b_2^*)$,

where b_1^* and b_2^* are any real number. The U1 and U2 sequences are depicted in Figure 8 in dark blue, the E sequences are depicted in dark purple, and the CC sequences are depicted in light purple. The E class tends to have both β_{1n} and β_{2n} converging at an equal or close to equal rate. The U1 and U2 classes tend to have β_{1n} and β_{2n} converging at an unequal rate. For U1, β_{1n} converges slower, and for U2, β_{2n} converges slower. The CC class is very close to the center in the sense that both β_{1n} and β_{2n} converges to zero at the $\sqrt{n}$ rate or faster. We can index these sequences by $g \in G_j$, which includes a local parameter (for U1, U2, or CC) and θ_0 for $j \in \{E, U1, U2, CC\}$. Theorem 2, applied to all classes of super-weak sequences, gives

$$\begin{pmatrix} \sqrt{n}(\hat{\psi}_n - \psi_n) \\ \hat{\pi}_n \end{pmatrix} \to_d \begin{pmatrix} Z_j^*(\pi_j^*) \\ \pi_j^* \end{pmatrix},$$

[28] See Section 3.2 for a definition of $\xi(\pi)$.

for $Z_j^*(\cdot)$, a Gaussian stochastic process over π, and π_j^*, a random variable whose distribution is characterized as the argmin of $\xi^*(\pi)$.[29] The continuous mapping theorem then implies that $W_n(\nu)$ converges to a nonstandard limit distribution, $F_j(g)$, for $g \in G_j$ and $j \in \{E, U1, U2, CC\}$.

4.2 Robust Critical Values

There are a variety of ways to combine the quantiles of these distributions to calculate robust critical values. The simplest way is to let

$$\hat{c}^{LF}_{\text{Robust}} = \max\left(\max_{j \in \{NS, EW, C, E, U1, U2, CC\}} \sup_{g \in G_j} F_{j,1-\alpha}(g), \chi^2_{d_r,1-\alpha}\right),$$

where $F_{j,1-\alpha}(g)$ denotes the $1-\alpha$ quantile of $F_j(g)$ and $\chi^2_{d_r,1-\alpha}$ denotes the $1-\alpha$ quantile of the $\chi^2_{d_r}$ distribution. This is the least favorable critical value in Andrews and Cheng (2012). This definition for $\hat{c}_{\text{Robust}}$ has two drawbacks. First, this supremum tends to be larger than necessary, which is conservative. A less conservative critical value would take a supremum over fewer distributions. Second, the distributions need to be simulated, and taking the supremum over many distributions is computationally burdensome. For both reasons, it is desirable to reduce the number of distributions. There are three ways this can be done.[30]

1. One can impose the null hypothesis on the values of θ_0 in the definition of G_j. Values of θ_0 that do not satisfy the null are unnecessary for controlling size. In this case, G_j depends on ν.

2. One can impose a consistent estimator for some parameters in θ_0. Since the quantiles are continuous with respect to these parameters, imposing a consistent estimator is equivalent to imposing the true value in the limit. The parameters in θ_0 that are consistently estimable are the ζ_0 parameters. In this case, G_j is random, denoted by $\hat{G}_j$.

3. One can define new test statistics, s_n, to consistently distinguish between some of the classes. To do this, one defines a sequence of partitions, $B_{m,n}$, for the range of s_n for $m = 1, ..., M$. Then, for each $m = 1, ..., M$, if $s_n \in B_{m,n}$, the critical value is defined by maximizing over $j \in J_m$, where J_m is a subset of $\{NS, EW, C, E, U1, U2, CC\}$.

[29] See Section 3.3 for a definition of $\xi^*(\pi)$.

[30] The first two correspond to the null-imposed and plug-in critical values of Andrews and Cheng (2012).

Using all these together, robust critical values can be defined by:

$$\hat{c}_{\text{Robust}} = \max\left(\sum_{m=1}^{M}\left(\mathbb{1}\{s_n \in B_{m,n}\} \sup_{j\in J_m} \sup_{g\in\hat{G}_n(\nu)} F_{j,1-\alpha}(g)\right), \chi^2_{d_r,1-\alpha}\right). \tag{4.1}$$

There are other approaches in the literature for calculating robust critical values out of classes of asymptotic distributions. Equation (4.1) corresponds to type 1 critical values in Andrews and Cheng (2012). Andrews and Cheng (2012) also define type 2 critical values, which use a softer transition from standard to robust critical values. Additionally, McCloskey (2012) constructs a confidence set for the local parameters and takes the supremum over local parameters in the confidence set, using the Bonferroni inequality to control size.

In Example 1, we can define

$$s_n = \begin{pmatrix} s_{1n} \\ s_{2n} \end{pmatrix} = \begin{pmatrix} \sqrt{n}\hat{\beta}_{1n}\hat{\text{Var}}(\hat{\beta}_{1n})^{-1/2} \\ \sqrt{n}\hat{\beta}_{2n}\hat{\text{Var}}(\hat{\beta}_{2n})^{-1/2} \end{pmatrix},$$

where $\text{Var}(\hat{\beta}_{in})$ is the asymptotic variance of $\hat{\beta}_{in}$ and $\hat{\text{Var}}(\hat{\beta}_{in})$ is an estimator of it.[31] Let $\bar{\kappa}_n \to \infty$ such that $n^{-1/2}\bar{\kappa}_n \to 0$. Then, we can define one partition to be $B_{1,n} = \{|s_{1n}| > \bar{\kappa}_n\}$ and $B_{2,n} = \{|s_{1n}| \leq \bar{\kappa}_n\}$. For these definitions, we notice that $|s_{1n}| > \bar{\kappa}_n$ with probability approaching 1 for weak sequences in the EW class, while $|s_{1n}| \leq \bar{\kappa}_n$ with probability approaching 1 for classes NS, U2, and CC. Thus, s_{1n} distinguishes between sequences and allows taking the supremum over fewer quantiles.

One can define finer partitions by using s_{1n} and s_{2n} together, which allows for even more refined critical values. For example, if we also assume that $n^{-1/4}\bar{\kappa}_n \to \infty$, then we can define

$$\begin{aligned} B_{1,n} &= \{|s_{1n}| > \bar{\kappa}_n, |s_{2n}| > \bar{\kappa}_n\} \\ B_{2,n} &= \{|s_{1n}| > \bar{\kappa}_n, |s_{2n}| \leq \bar{\kappa}_n\} \\ B_{3,n} &= \{|s_{1n}| \leq \bar{\kappa}_n, |s_{2n}| > \bar{\kappa}_n\} \\ B_{4,n} &= \{|s_{1n}| \leq \bar{\kappa}_n, |s_{2n}| \leq \bar{\kappa}_n\}. \end{aligned}$$

Then, $B_{1,n}$ only occurs with a positive probability in the limit for strong and semi-strong sequences. $B_{2,n}$ occurs with a positive probability in the limit for C, U1, EW, strong, and semi-strong sequences. $B_{3,n}$ occurs with a positive probability in the limit for C, U2, NS, strong, and semi-strong sequences. $B_{4,n}$ occurs with a positive probability in the limit for C, E, U1, U2, CC, and semi-strong sequences. This justifies defining $J_1 = \emptyset$, $J_2 = \{C, U1, EW\}$, $J_3 = \{C, U2, NS\}$, and $J_4 = \{C, E, U1, U2, CC\}$. Equation (4.1) then defines $\hat{c}_{\text{Robust}}$ for Example 1.

[31] Formulas for $\hat{\text{Var}}(\hat{\beta}_{1n})$ and $\hat{\text{Var}}(\hat{\beta}_{2n})$ are given in Supplemental Materials 2.

Let

$$t_n = \sqrt{n}|\hat{\beta}_{1n}\hat{\beta}_{2n}| \left(\hat{\beta}_{2n}^2 \hat{\text{Var}}(\hat{\beta}_{1n}) + 2\hat{\beta}_{1n}\hat{\beta}_{2n}\hat{\text{Cov}}(\hat{\beta}_{1n}, \hat{\beta}_{2n}) + \hat{\beta}_{1n}^2 \hat{\text{Var}}(\hat{\beta}_{2n}) \right)^{-1/2},$$

and κ_n be a sequence of constants such that $n^{-1/4}\kappa_n \to \infty$ and $n^{-1/2}\kappa_n \to 0$.[32] t_n is the absolute value of the t test statistic for testing the hypothesis $\beta_1\beta_2 = 0$.

The fact that κ_n must diverge at least as fast as $n^{1/4}$ is strange. The reason for this is the fact that there exist weak sequences such that t_n diverges. This divergence comes from the fact that the denominator converges to zero when both β_{1n} and β_{2n} converge to zero. The divergence of t_n is a manifestation of the degeneracy of the first derivative of the hypothesis $\beta_1\beta_2 = 0$. For all weak and super-weak sequences, the fastest rate at which t_n diverges is $n^{1/4}$. For this reason, any κ_n that satisfies the above conditions consistently distinguishs strong sequences from weak and super-weak sequences.

The robust inference procedure is then defined to reject H_0 if either $t_n > \kappa_n$ and $W_n(\nu) > \chi^2_{d_r,1-\alpha}$ or if $t_n \leq \kappa_n$ and $W_n(\nu) > \hat{c}_{\text{Robust}}$. Theorem 5, in Supplemental Materials 2, states that this robust inference procedure controls size for Example 1, whether or not the rank condition holds.

5 Robust Inference with Two Factors

This section provides a method for robust inference in Example 2, a factor model with two factors. Robust inference follows the same two step approach as Section 4. Consider testing a null hypothesis, $H_0 : r(\theta) = \nu$, using the Wald test statistic, $W_n(\nu)$. The first step tests the rank condition. The second step uses standard chi-squared critical values or robust critical values depending on the outcome of the first step.

This section defines Example 2, reparameterizes the model into the setup of Section 3, divides sequences of parameters into classes, characterizes the asymptotic distribution of the estimator and test statistic within each class, and describes how to calculate robust critical values.

5.1 Example 2: Two Factors

Consider the case where there are only two factors and five observed variables. An extension to more observed variables is handled in Appendix B. In this case, the matrix of

[32] $\hat{\text{Cov}}(\hat{\beta}_{1n}, \hat{\beta}_{2n})$ is an estimator of the asymptotic covariance between $\hat{\beta}_{1n}$ and $\hat{\beta}_{2n}$.

factor loadings is given by:

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \end{bmatrix},$$

where the first two rows represent the normalization. The other parameters in the model are σ_1^2, the variance of the first factor, σ_2^2, the variance of the second factor, σ_{12}, the covariance between the two factors, and ϕ_j for $j = 1, 2, ..., 5$, the variances of the error terms. For this example, we assume that the first factor is strong, so that $\lambda_{11} \neq 0$ and $\lambda_{21} \neq 0$. All the other lambdas are allowed to take any value.

In this case, there are three ways the rank condition can fail.

1. Weak Measurement: This occurs when both λ_{31} and λ_{32} are zero. In this case, the fifth observed variable is uncorrelated with both factors. Essentially, there are only four observed variables, which is insufficient to identify two factors.

2. Weak Factor: This occurs when two of λ_{12}, λ_{22}, or λ_{32} are zero. In this case, the second factor does not have three nonzero factor loadings, and so it cannot be identified. An extreme case of this is when all three lambdas are zero. Then, the second factor does not explain any of the covariation in the data, so the model essentially has only one factor. In this sense, Example 2 nests the one factor model and allows for an unknown number of factors (either one or two).

3. Entangled Factors: This occurs when the lambdas for the second factor are a scalar multiple of the lambdas for the first factor. That is, there exists an $\alpha \in \mathbb{R}$ such that

   $$\begin{pmatrix} \lambda_{12} \\ \lambda_{22} \\ \lambda_{32} \end{pmatrix} = \alpha \begin{pmatrix} \lambda_{11} \\ \lambda_{21} \\ \lambda_{31} \end{pmatrix}.$$

   Intuitively, this occurs when observed variables numbered 3 through 5 depend only on a particular linear combination of the two factors, so that the distribution of the two factors cannot be separately identified. This source of identification failure is another way to nest the one factor model because when $\alpha = 0$, there is essentially only one factor. This source for identification failure is important for the empirical application in Section 6.

5.2 Reparameterization

This section describes a reparameterization that translates Example 2 into the setup of Definition Parameter Spaces and Definition Objective Functions.

The reduced form parameters in this model are the elements of the covariance matrix of the observed variables, equal to $Cov(X_i) = \Omega(\theta) = \Lambda\Sigma\Lambda' + \Phi$. The following equations define the reparameterization.

$$\begin{aligned}
\beta_{11} &= \lambda_{11}\sigma_1^2 + \lambda_{12}\sigma_{12} \\
\beta_{12} &= \lambda_{11}\sigma_{12} + \lambda_{12}\sigma_2^2 \\
\beta_{21} &= \lambda_{21}\sigma_1^2 + \lambda_{22}\sigma_{12} \\
\beta_{22} &= \lambda_{21}\sigma_{12} + \lambda_{22}\sigma_2^2 \\
\beta_{31} &= \lambda_{31}\sigma_1^2 + \lambda_{32}\sigma_{12} \\
\beta_{32} &= \lambda_{31}\sigma_{12} + \lambda_{32}\sigma_2^2 \\
\zeta_1 &= \sigma_1^2 + \phi_1 \\
\zeta_2 &= \sigma_2^2 + \phi_2 \\
\zeta_3 &= \begin{pmatrix} \lambda_{11} \\ \lambda_{12} \end{pmatrix}' \Sigma \begin{pmatrix} \lambda_{11} \\ \lambda_{12} \end{pmatrix} + \phi_3 \\
\zeta_4 &= \begin{pmatrix} \lambda_{21} \\ \lambda_{22} \end{pmatrix}' \Sigma \begin{pmatrix} \lambda_{21} \\ \lambda_{22} \end{pmatrix} + \phi_4 \\
\zeta_5 &= \begin{pmatrix} \lambda_{31} \\ \lambda_{32} \end{pmatrix}' \Sigma \begin{pmatrix} \lambda_{31} \\ \lambda_{32} \end{pmatrix} + \phi_5 \\
\rho &= \begin{pmatrix} \lambda_{11} \\ \lambda_{12} \end{pmatrix}' \Sigma \begin{pmatrix} \lambda_{21} \\ \lambda_{22} \end{pmatrix} \\
\pi &= \sigma_2^2.
\end{aligned}$$

This reparameterization allows us to write the covariance of the observed variables as[33]

$$\Omega(\theta) = \begin{bmatrix} \zeta_1 & \sigma_{12} & \beta_{11} & \beta_{21} & \beta_{31} \\ \sigma_{12} & \zeta_2 & \beta_{12} & \beta_{22} & \beta_{32} \\ \beta_{11} & \beta_{12} & \zeta_3 & \rho & h_1(\psi,\pi) + \rho\beta_{31}\beta_{21}^{-1} \\ \beta_{21} & \beta_{22} & \rho & \zeta_4 & h_2(\psi,\pi) + \rho\beta_{31}\beta_{11}^{-1} \\ \beta_{31} & \beta_{32} & h_1(\psi,\pi) + \rho\beta_{31}\beta_{21}^{-1} & h_2(\psi,\pi) + \rho\beta_{31}\beta_{11}^{-1} & \zeta_5 \end{bmatrix},$$

where $\beta = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \beta_{31}, \beta_{32})$, $\zeta = (\zeta_1, \zeta_2, \zeta_3, \zeta_4, \zeta_5)$, $\psi = (\beta, \zeta, \sigma_{12}, \rho)$, and

$$\begin{aligned} h_1(\psi,\pi) &= \frac{\beta_{12}(\beta_{32}\beta_{21} - \beta_{31}\beta_{22})}{\pi\beta_{21} - \sigma_{12}\beta_{22}} + \rho\frac{\beta_{31}\pi - \beta_{32}\sigma_{12}}{\pi\beta_{21} - \sigma_{12}\beta_{22}} - \rho\beta_{31}\beta_{21}^{-1} \\ h_2(\psi,\pi) &= \frac{\beta_{22}(\beta_{32}\beta_{11} - \beta_{31}\beta_{12})}{\pi\beta_{11} - \sigma_{12}\beta_{12}} + \rho\frac{\beta_{31}\pi - \beta_{32}\sigma_{12}}{\pi\beta_{11} - \sigma_{12}\beta_{12}} - \rho\beta_{31}\beta_{11}^{-1}. \end{aligned}$$

One can verify that the ψ and h parameters are identified, while the π parameter is identified if and only if

$$\beta_{32}\beta_{21} - \beta_{31}\beta_{22} \neq 0 \text{ or } \beta_{32}\beta_{11} - \beta_{31}\beta_{12} \neq 0.$$

This is a version of the rank condition for Example 2, translated into the new parameters. Notice that both conditions represent determinants on a matrix of β's. If both conditions fail, then the rank condition fails, and the model has a weak fifth measurement, a weak second factor, or the two factors are entangled.

When the rank condition fails, π is still partially identified by bounds. These bounds come from nonnegativity of the variance parameters and the fact that the variance matrix of the factors is positive definite. These bounds become intractable under the reparameterization. Fortunately, we do not have to keep track of them explicitly, since any identified set with a continuous boundary satisfies Assumption HC. Supplemental Materials 3 gives more details on the definition of this model, including the specifications of the parameter space and likelihood.

5.3 Asymptotic Distributions

Robust critical values depend on the asymptotic distribution of $W_n(\nu)$ along sequences of parameters converging to points of rank condition failure. This section divides such sequences into classes and characterizes the asymptotic distribution of the estimator and the Wald test

[33] Even though λ_{11} and λ_{21} are assumed to be nonzero, it is technically possible that β_{11} or β_{21} could still be zero. In this example, I assume that $\beta_{11} \neq 0$ and $\beta_{21} \neq 0$, which says that the first and second non-normalized observed variables, in addition to having nonzero factor loadings on the first factor, have a nonzero covariance with the first factor. A slightly different formula for $h(\psi, \pi)$ works when $\beta_{11} = 0$ or $\beta_{21} = 0$.

statistic along sequences in each class.

By Assumption h1, the classes are defined based on the rate at which $h(\psi_n, \pi) = (h_1(\psi_n, \pi), h_2(\psi_n, \pi))$ converges to $(0, 0)$. This only depends on the rates at which $a_{1n} = \beta_{32n}\beta_{21n} - \beta_{31n}\beta_{22n}$ and $a_{2n} = \beta_{32n}\beta_{11n} - \beta_{31n}\beta_{12n}$ converge to zero.[34]

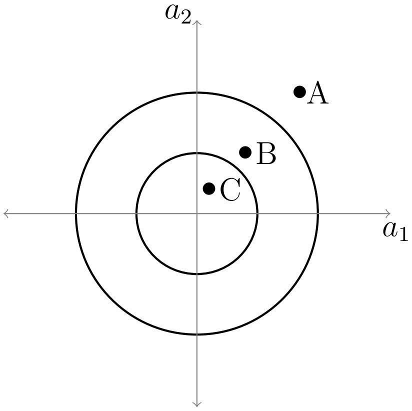

Figure 9

These classes are illustrated using a heuristic sketch, depicted in Figure 9. A point outside both circles, for example point A, depicts a sequence, θ_n, that converges to a limit, θ_0, for which $a_{10} \neq 0$ or $a_{20} \neq 0$. A point between the circles, for example point B, depicts a sequence that converges to a limit, θ_0, such that both $a_{10} = 0$ and $a_{20} = 0$, but the rate of convergence for one or both of them is slower than $n^{-1/2}$. A point inside both circles, for example point C, depicts a sequence for which (a_{1n}, a_{2n}) converges to $(0, 0)$ at an $n^{-1/2}$ rate or faster.

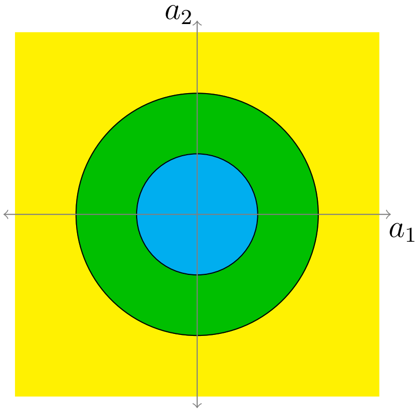

Figure 10

First, we can define the class of strong sequences, θ_n, as those that converge to a limit, θ_0, such that either $a_{10} \neq 0$ or $a_{20} \neq 0$. These sequences are depicted in Figure 10 in yellow. Theorem 4, in Supplemental Materials 1, applied to strong sequences, gives an asymptotic normal distribution for $\hat{\theta}_n$.[35] The continuous mapping theorem implies that $W_n(\nu)$ converges to a limiting $\chi^2_{d_r}$ distribution.

The class of semi-strong sequences is characterized by $h(\psi_n, \pi)$ converging to zero at a rate slower than $n^{-1/2}$. This occurs when $(a_{10}, a_{20}) = (0, 0)$ and $\sqrt{n}||(a_{1n}, a_{2n})|| \to \infty$. These sequences are depicted in Figure 10 in green. Theorem 3, in Supplemental Materials 1, applied to semi-strong sequences, gives

$$\begin{pmatrix} \sqrt{n}(\hat{\psi}_n - \psi_n) \\ \sqrt{n}a_n^{-1}(\hat{\pi}_n - \pi_n) \end{pmatrix} \to_d Z_S,$$

where $a_n \to \infty$ and Z_S is a normal random variable. a_n diminishes the rate of convergence of $\hat{\pi}_n$. The continuous mapping theorem implies that $W_n(\nu)$ converges to a limiting $\chi^2_{d_r}$ distribution.

[34] Notice that $h_1(\psi_n, \pi) = \dfrac{\beta_{32n}\beta_{21n} - \beta_{31n}\beta_{22n}}{\pi\beta_{21n} - \sigma_{12n}\beta_{22n}}\left(\beta_{12n} - \dfrac{\rho_n\sigma_{12n}}{\beta_{21n}}\right)$, which goes to zero as a_{1n} goes to zero. A similar calculation holds for $h_2(\psi_n, \pi)$.

[35] Supplemental Materials 3 verifies the assumptions of Theorem 4. More generally, Supplemental Materials 3 contains details and proofs for the assertions in this section (Section 5).

The class of weak sequences is characterized by $h(\psi_n, \pi)$ converging to zero at an $n^{-1/2}$ rate or faster. This occurs when $(a_{10}, a_{20}) = (0, 0)$ and $\sqrt{n}(a_{1n}, a_{2n}) \to (b_1, b_2)$. These sequences are depicted in Figure 10 in light blue. Theorem 1, applied to weak sequences, gives

$$\begin{pmatrix} \sqrt{n}(\hat{\psi}_n - \psi_n) \\ \hat{\pi}_n \end{pmatrix} \to_d \begin{pmatrix} Z_W(\pi_W^*) \\ \pi_W^* \end{pmatrix},$$

where $Z_W(\cdot)$ is a Gaussian stochastic process over π, and π_W^* is a random variable whose distribution is characterized as the argmin of a stochastic process. This distribution is indexed by $(b_1, b_2) \in \mathbb{R}^2$ and θ_0. We can combine these into a single index $g \in G_W$. The continuous mapping theorem implies $W_n(\nu)$ converges to a nonstandard limit distribution, $F_W(g)$ for $g \in G_W$.

5.4 Robust Critical Values

Similar to Example 1, there are a variety of ways to combine the quantiles of these distributions to calculate robust critical values. The simplest way is to let

$$\hat{c}_{\text{Robust}}^{LF} = \max\left(\sup_{g \in G_W} F_{W,1-\alpha}(g), \chi^2_{d_r,1-\alpha}\right),$$

where $F_{W,1-\alpha}(g)$ denotes the $1-\alpha$ quantile of $F_W(g)$. This definition suffers from the same two drawbacks as Example 1. Thus, it is desirable to reduce the number of distributions. Analogous to Example 1, one can impose the null hypothesis on the values of θ_0 in the definition of G_W. Also, one can plug in consistent estimators for parameters in θ_0.[36] In Example 2, the consistently estimable parameters are β_0, ζ_0, ρ_0, and σ_{120}. However, because there is only one class of sequences, there is no way to consistently distinguish between different classes of sequences. After imposing the null and plugging in consistent estimators, the robust critical values are defined to be:

$$\hat{c}_{\text{Robust}} = \max\left(\sup_{g \in \hat{G}_W(\nu)} F_{W,1-\alpha}(g), \chi^2_{d_r,1-\alpha}\right).$$ [37]

Let

$$t_n = \left(\frac{n}{2}\gamma(\hat{\beta}_n)' \left(B(\hat{\beta}_n)\hat{\text{Var}}(\hat{\theta}_n)B(\hat{\beta}_n)'\right)^{-1} \gamma(\hat{\beta}_n)\right)^{1/2},$$

[36] These critical values correspond to the null-imposed and plug-in least favorable critical values in Andrews and Cheng (2012), respectively.

[37] Similar to Example 1, there are other ways in the literature to define robust critical values, including type 2 critical values in Andrews and Cheng (2012) and Bonferroni based critical values in McCloskey (2012).

where

$$\begin{aligned}\gamma(\beta) &= \begin{pmatrix} \beta_{21}\beta_{32} - \beta_{22}\beta_{31} \\ \beta_{11}\beta_{32} - \beta_{12}\beta_{31} \end{pmatrix}, \\ B(\beta) &= \begin{bmatrix} 0 & 0 & \beta_{32} & -\beta_{31} & -\beta_{22} & \beta_{21} & 0_{1\times 8} \\ \beta_{32} & -\beta_{31} & 0 & 0 & -\beta_{12} & \beta_{11} & 0_{1\times 8} \end{bmatrix},\end{aligned}$$

and $\hat{\text{Var}}(\hat{\theta}_n)$ is an estimator of the asymptotic variance of $\hat{\theta}_n$ that is consistent under strong and semi-strong sequences.[38] t_n is related to the Wald test statistic for testing the rank condition, formulated as the hypothesis, $\gamma(\beta) = 0$. $B(\beta)$ is the derivative of $\gamma(\beta)$ with respect to $\psi = (\beta, \zeta, \rho, \sigma_{12})$. Also let $\kappa_n \to \infty$ such that $n^{-1/2}\kappa_n \to 0$. For Example 2, the restriction that $n^{-1/4}\kappa_n \to \infty$ is not needed. This is because the hypothesis $\gamma(\beta) = 0$ has no degeneracy. Thus, any κ_n diverging at a rate slower than $n^{1/2}$ is sufficient to consistently distinguish between strong and weak sequences.

The robust inference procedure is defined to reject H_0 if either $t_n > \kappa_n$ and $W_n(\nu) > \chi^2_{d_r,1-\alpha}$ or if $t_n \leq \kappa_n$ and $W_n(\nu) > \hat{c}_{\text{Robust}}$. Theorem 6, in Supplemental Materials 3, states that this robust inference procedure controls size for Example 2, whether or not the factors are identified.

6 Robust Inference for the Production of Cognitive and Noncognitive Skills in Children

Cunha and Heckman (2008) and Cunha, Heckman, and Schennach (2010) estimate the production of skills in children as a function of parental investments. They model parental investments and child skills as unobserved factors common to a variety of measures of home environment, parental activities, as well as cognitive and personality test scores. When they formulate the model, they include two types of skills, cognitive and noncognitive skills, with two types of parental investments. When taking the model to the data, they encounter the problem that two types of parental investments are not identified.

> "In practice, we cannot empirically distinguish investments in cognitive skills from investments in noncognitive skills. Accordingly, we assume investment in period t is the same for both skills, although it may have different effects on those skills. Thus we assume $I_{C,t} = I_{N,t}$ and define it as I_t."[39]

Assuming only a one dimensional investment factor eliminates important questions about

[38] A formula for $\hat{\text{Var}}(\hat{\theta}_n)$ is given in Supplemental Materials 3.

[39] Cunha, Heckman, and Schennach (2010), page 904.

Category	Name of Observed Variable
Cognitive Normalization:	Number of books the child has
Noncognitive Normalization:	How often the child sees family friends
Cognitive Investments:	How often the child is taken to a museum How often mom reads to the child
Additional Variables:	How often the child eats with mom/dad Whether the family receives a daily newspaper Whether the child is taken to musical performances

Table 2: Observed variables for the model of parental investments in 6-9 year-old children.

the relationship between investment in cognitive skills and investment in noncognitive skills and the effects these investments have on the development of cognitive and noncognitive skills for different stages of development. Instead of assuming one parental investment factor, this paper proposes using robust inference in the model for two factors. With this approach, two investment factors do not need to be identified. Instead, a first stage test of the rank condition determines whether one uses standard or robust critical values.

6.1 The Distribution of Parental Investments

A factor model for parental investments assumes that the covariation between a variety of observed variables related to parental interactions with children and home environment can be summarized in two dimensions of parental investment, investment in cognitive skills and investment in noncognitive skills. The model is specified using data on parental interactions and the home environment that are included in the CNLSY/79 Home Observation of the Environment - Short Form (HOME-SF). These variables are a subset of the ones used by Cunha, Heckman, and Schennach (2010).

For example, Table 2 lists the observed variables used for the 6-9 year-old specification.[40] First notice that there are seven variables used, which is larger than the five variables for Example 2. Appendix B describes how to extend Example 2 to more observed variables. Next, notice that the variables need to be categorized. The number of books is chosen to be the normalization for investment in cognitive skills. This assumes that this variable constitutes an investment in cognitive skills and does not constitute an investment in noncognitive skills. This is reasonable because a parent purchasing a child a book is an indicator that the parent is investing in the child's knowledge or reading skills, which are types of cognitive skills, and does not indicate the the parent is investing in any noncognitive skills of the child, such as

[40] A list of the observed variables that are used in other age specifications can be found in Section A. Full documentation can be found in Center for Human Resource Research (2006).

social or behavioral skills. Analogously, family friends is chosen to be the normalization for investment in noncognitive skills. This assumes that this variable constitutes an investment in noncognitive skills and does not constitute an investment in cognitive skills. This is reasonable because a parent inviting a family member or a family friend over to visit with the child is an indicator that the parent is investing in the child's social or conversational skills, which are types of noncognitive skills, and does not indicate that the parent is investing in any cognitive skills of the child.[41] Also, the specification of Example 2 assumes that there are two more variables which constitute investments in one of the factors. For this specification, those variables are taken to be museum and reads, which constitute investments in cognitive skills. This is reasonable because taking a child to a museum and reading to a child is an indicator that the parent is investing in the knowledge or reading/language skills of the child, which are types of cognitive skills. The additional variables are allowed to be arbitrarily related or unrelated to investments in cognitive or noncognitive skills.

These assumptions are very weak for this factor model, especially compared to the assumptions required for identification. Identification would require assuming that at least two of cognitive investment variables or additional variables constitute investments in noncognitive skills (at least one of which is an additional variable). Identification would also require that the linear combination of investment in cognitive and noncognitive skills is different for at least two cognitive investment variables or additional variables. These assumptions are unwarranted in this example because we want to allow for the fact that all of the cognitive investments and additional variables may be completely unrelated to investment in noncognitive skills. The robust inference procedure for Example 2 does not require identification, so these assumptions are not needed.

Table 3 gives estimates for the parameters in the distribution of the investment factors, as well as standard and robust confidence intervals. σ_1^2 is the variance of investment in cognitive skills, σ_2^2 is the variance of investment in noncognitive skills, and σ_{12} is the covariance. t_n is the value of the first stage test statistic for the rank condition and κ_n is a drifting threshold.[42] For $t_n > \kappa_n$, the model is probably identified and standard critical values can be used to construct confidence intervals. For $t_n \leq \kappa_n$, the model may or may not be identified and robust critical values must be used.

Table 3 shows that three of the specifications are probably identified: 0-2 year-old, 3 -5 year-old, and 10-12 year-old. For 6-9 year-old, the first stage test statistic is not large enough, so robust critical values must be used. Consequently, the robust confidence intervals for the identified specifications agree with the standard confidence intervals, while the robust

[41] An exception to this could be if the visitor has special knowledge that the parents want to impart to the child, such as a tutor or someone who speaks a different language. This normalization assumes that the presence of this exception is negligible in the data.

[42] For the empirical specification, κ_n is chosen to be $\sqrt{\log(n)}$, analogous to BIC.

Age	Type	σ_1^2	σ_{12}	σ_2^2	t_n	κ_n
0-2 Years	Estimate	0.62	0.32	0.38	3.2	2.7
	Standard	[0.41 , 0.83]	[0.24 , 0.39]	[0.17 , 0.58]		
	Robust	[0.41 , 0.83]	[0.24 , 0.39]	[0.17 , 0.58]		
3-5 Years	Estimate	0.028	0.018	0.110	6.0	2.8
	Standard	[0 , 0.107]	[-0.011 , 0.047]	[0.032 , 0.189]		
	Robust	[0 , 0.107]	[-0.011 , 0.047]	[0.032 , 0.189]		
6-9 Years	Estimate	0.007	0.014	0.026	1.6	2.8
	Standard	[0 , 0.042]	[-0.005 , 0.033]	[0 , 0.061]		
	Robust	[0 , 0.574]	[-0.005 , 0.033]	[0 , 0.481]		
10-12 Years	Estimate	0.015	0.122	11.05	3.5	2.8
	Standard	[0 , 0.049]	[0.089 , 0.149]	[11.01 , 11.09]		
	Robust	[0 , 0.049]	[0.089 , 0.149]	[11.01 , 11.09]		

Table 3: Parameter estimates and standard and robust confidence intervals for the variance and covariance of parental investment in cognitive and noncognitive skills.

confidence intervals for 6-9 year-old are larger. The robust confidence intervals for σ_1^2 and σ_2^2 are significantly larger than their standard confidence intervals, while the robust confidence interval for σ_{12} is the same as the standard confidence interval. This is because σ_{12} is identified, and the robust critical value for σ_{12} coincides with the chi-squared critical value, while the robust critical values for σ_1^2 and σ_2^2 are larger than the chi-squared critical value.[43]

Finally, some of the confidence intervals for the identified specifications are quite wide, despite being classified as identified. For example, the right end point of the confidence interval for σ_1^2 in the 3-5 year-old specification is more than three times the value of the estimate. This shows that wide confidence intervals are not reliable indicators of identification status. A confidence interval also could be wide because the data is noisy.

What is different about the 6-9 year old specification that causes it to be unidentified, while the other specifications are identified? Each cognitive investment variable and additional variable constitutes an investment in a linear combination of cognitive and noncognitive skills. The problem is that each of these variables constitutes an investment in the same linear combination of cognitive and noncognitive skills. This is detected by estimating the covariances between the observed variables and the parental investment factors, displayed in Table 4. The right column is approximately a scalar multiple of the left column.[44] This

[43] For identified parameters, robust critical values sometimes coincide with chi-squared critical values, but not always.

[44] The ratio between the right and left column is always between 1.4 and 3.3, a relatively small range considering statistical error.

	Cognitive	Noncognitive
Museum	0.026	0.085
Music	0.029	0.068
Reads	0.044	0.078
Eats	0.025	0.034
Newspaper	0.012	0.022

Table 4: Estimated covariances between observed variables and investments in cognitive and noncognitive skills of 6-9 year-old children.

means that, at the estimated value of the parameters, the factors are close to being entangled. The first stage test of the rank condition cannot reject the null hypothesis that the rank condition fails, and the model is close to unidentified.

6.2 Estimating the Production Function

This section describes how to perform robust inference for parameters in a production function where the inputs to production are unobserved factors. Robust inference for these parameters involves, first, specifying a robust inference procedure for parameters, θ, that index the first stage factor model, and second, writing the parameters in the production function as a function of the first stage parameters, $r(\theta)$. This section discusses both linear and nonlinear production functions.

6.2.1 Linear Production Functions

Cunha and Heckman (2008) combine the factor model for parental investments with factors for the distribution of skills, both cognitive and noncognitive, over time as well as a factor for maternal abilities. They specify a linear production function for cognitive and noncognitive skills. Allowing for two parental investments, the linear production function is

$$\begin{aligned}
ln(S_{C,t+1}) &= \gamma_{C1} ln(S_{C,t}) + \gamma_{C2} ln(S_{N,t}) + \gamma_{C3} ln(I_{C,t}) + \gamma_{C4} ln(M_A) + \gamma_{C5} ln(M_E) + \epsilon_{C,t+1} \\
ln(S_{N,t+1}) &= \gamma_{N1} ln(S_{C,t}) + \gamma_{N2} ln(S_{N,t}) + \gamma_{N3} ln(I_{N,t}) + \gamma_{N4} ln(M_A) + \gamma_{N5} ln(M_E) + \epsilon_{N,t+1},
\end{aligned}$$

where $S_{C,t}$ is child cognitive skills, $S_{N,t}$ is child noncognitive skills, $I_{C,t}$ is parental investment in cognitive skills, $I_{N,t}$ is parental investment in noncognitive skills, M_A is maternal abilities (an unobserved factor), and M_E is mother's education (an observed variable).

In the Cunha and Heckman (2008) specification, expanded by a second investment factor, the first step comes from applying the robust inference procedure of Section 3 to a model with more factors. In this case the rank condition and corresponding identification status is more complicated than the rank condition for a model with two factors. Appendix C discusses extending the robust inference of Examples 1 and 2 to more factors.

For the second step, let $RHS_{C,t}$ and $RHS_{N,t}$ collect the right hand side variables (as 5×1 vectors) for the cognitive and noncognitive equations, respectively. Then, let

$$\text{Var}\begin{pmatrix} S_{C,t+1} \\ RHS_{C,t} \end{pmatrix} = \begin{bmatrix} \sigma^2_{C,t+1} & \sigma'_{C,t+1,t} \\ \sigma_{C,t+1,t} & \Sigma_{C,t} \end{bmatrix} \text{ and } \text{Var}\begin{pmatrix} S_{N,t+1} \\ RHS_{N,t} \end{pmatrix} = \begin{bmatrix} \sigma^2_{N,t+1} & \sigma'_{N,t+1,t} \\ \sigma_{N,t+1,t} & \Sigma_{N,t} \end{bmatrix}$$

denote the covariance matrices of the variables in the two production functions, where $\sigma^2_{C,t+1}$ is the variance of $S_{C,t+1}$, $\sigma^2_{N,t+1}$ is the variance of $S_{N,t+1}$, $\Sigma_{C,t}$ is the 5×5 variance matrix of $RHS_{C,t}$, $\Sigma_{N,t}$ is the 5×5 variance matrix of $RHS_{N,t}$, $\sigma_{C,t+1,t}$ is the 5×1 vector of covariances of $RHS_{C,t}$ with $S_{C,t+1}$, and $\sigma_{N,t+1,t}$ is the 5×1 vector of covariances of $RHS_{N,t}$ with $S_{N,t+1}$. These parameters constitute part of the vector of first stage parameters, θ. We can write the coefficients in the production function as

$$\begin{aligned} \gamma_C &= \Sigma^{-1}_{C,t}\sigma_{C,t+1,t} \\ \gamma_N &= \Sigma^{-1}_{N,t}\sigma_{N,t+1,t}, \end{aligned}$$

where $\gamma_C = (\gamma_{C1}, \gamma_{C2}, \gamma_{C3}, \gamma_{C4}, \gamma_{C5})'$ and $\gamma_N = (\gamma_{N1}, \gamma_{N2}, \gamma_{N3}, \gamma_{N4}, \gamma_{N5})'$. This defines $r(\theta)$. Constructing the Wald test statistic for the hypothesis $H_0 : r(\theta) = \nu$ and using robust critical values gives robust inference.

6.2.2 Nonlinear Production Function

Cunha, Heckman, and Schennach (2010) specify a nonlinear production function with one parental investment factor. When expanded to allow for two parental investment factors, the production function is

$$\begin{aligned} S_{C,t+1} &= [\gamma_{C,1}S^{\phi_C}_{C,t} + \gamma_{C,2}S^{\phi_C}_{N,t} + \gamma_{C,3}I^{\phi_C}_{C,t} + \gamma_{C,4}M^{\phi_C}_{C} + \gamma_{C,5}M^{\phi_C}_{N}]^{1/\phi_C}e^{\eta_{C,t+1}} \\ S_{N,t+1} &= [\gamma_{N,1}S^{\phi_N}_{C,t} + \gamma_{N,2}S^{\phi_N}_{N,t} + \gamma_{N,3}I^{\phi_N}_{N,t} + \gamma_{N,4}M^{\phi_N}_{C} + \gamma_{N,5}M^{\phi_N}_{N}]^{1/\phi_N}e^{\eta_{N,t+1}}, \end{aligned}$$

where M_C is maternal cognitive skills and M_N is maternal noncognitive skills, which are both unobserved factors.

The first step, specifying a robust inference procedure for parameters, θ, that index the first stage factor model, is in principle the same as for the linear production function. However, a potential difference comes from the fact that the parameters in the nonlinear regression depend on properties of the joint distribution of the factors beyond second moments. For this reason, it is important to allow for more flexibility in the joint distribution of the factors. Cunha, Heckman, and Schennach (2010) include specifications where the parametric form for the distribution of the factors is a mixture of normals. This complicates the identification characterization both because mixture models may not be identified and because the rank condition on the factor loadings has a different effect on the identification of the parameters in the distribution of the factors.[45] Robust inference for this specification requires reparameterizing the model into the framework of Section 3.1 and applying Theorems 1 and 2 to drifting sequences of parameters.

The second step is to write the parameters in the production function as a function of the parameters in the distribution of the factors. In this case, there is no closed form solution, but this can be calculated by simulation as in Attanasio, Meghir, and Nix (2015).

7 Conclusion

Identification in factor models can be characterized by a rank condition on the factor loadings. The literature on identification in factor models has focused on providing sufficient conditions for the rank condition to hold generically over the parameter space. However, this leads to poor inference results when the rank condition fails or is close to failing. This paper provides a method for robust inference based on calculating robust critical values that control rejection probabilities along sequences of parameters converging to points of rank condition failure. Within a class of models that are doubly parameterized by structural and reduced form parameters, two new theorems provide limit theory for weak and super-weak sequences of parameters. These theorems allow for nondifferentiability of the boundary of the identified set and degeneracy in the limit of the objective function.

Explicit procedures for robust inference are provided for two examples. The first example has one factor, that may be weak, and the second example has two factors, that may be entangled. The entanglement of the second example occurs in an empirical application of estimating the distribution of parental investments in the cognitive and noncognitive skills of children. In one of the age categories, the parental investment factors are entangled, and therefore not identified, while in the other three age categories, the parental investment factors are identified. This illustrates that the robust inference procedure is capable of distinguishing between specifications that are identified and those that are unidentified because of the rank condition.

[45] For a description of identification in mixture models, see Chen, Ponomareva, and Tamer (2014) and the references therein.

References

[1] Agostinelli, F., and M. Wiswall (2016a): "Estimating the Technology of Children's Skill Formation," NBER Working Paper 22442.

[2] Agostinelli, F., and M. Wiswall (2016b): "Identification of Dynamic Latent Factor Models: The Implications of Re-Normalization in a Model of Child Development," NBER Working Paper 22441.

[3] Almlund, M., A. Duckworth, J. Heckman, and T. Kautz (2011): "Personality Psychology and Economics," NBER Working Paper 16822.

[4] Anderson, T. W., and H. Rubin (1956): "Statistical Inference in Factor Analysis," *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, 5, 111-150.

[5] Andrews, I. (2016): "Conditional Linear Combination Tests for Weakly Identified Models," *Econometrica*, forthcoming.

[6] Andrews, D. W. K., and X. Cheng (2012): "Estimation and Inference with Weak, Semi-strong and Strong Identification," *Econometrica*, 80, 2153-2211.

[7] Andrews, D. W. K., and X. Cheng (2013): "Maximum Likelihood Estimation and Uniform Inference with Sporadic Identification Failure," *Journal of Econometrics*, 173, 36-56.

[8] Andrews, D. W. K., and X. Cheng (2014): "GMM Estimation and Uniform Subvector Inference with Possible Identification Failure," *Econometric Theory*, 30, 287-333.

[9] Andrews, D. W. K., and P. Guggenberger (2015): "Identification- and Singularity-Robust Inference for Moment Condition Models," Cowles Foundation Discussion Paper 1978.

[10] Andrews, D. W. K., and P. Guggenberger (2016): "Asymptotic Size of Kleibergen's LM and Conditional LR Tests for Moment Condition Models," *Econometric Theory*, forthcoming.

[11] Andrews, I., and A. Mikusheva (2016a): "A Geometric Approach to Nonlinear Econometric Models," *Econometrica*, 84, 1249-1264.

[12] Andrews, I., and A. Mikusheva (2016b): "Conditional Inference with a Functional Nuisance Parameter," *Econometrica*, 84, 1571-1612.

[13] Attanasio, O., S. Cattan, E. Fitzsimons, C. Meghir, and M. Rubio-Codina (2015): "Estimating the Production Function for Human Capital: Results from a Randomized Control Trial in Colombia," NBER Working Paper 20965.

[14] Attanasio, O., C. Meghir, and E. Nix (2015): "Human Capital Development and Parental Investment in India," NBER Working Paper 21740.

[15] Bai, J., and K. Li (2012): "Statistical Analysis of Factor Models of High Dimension," *The Annals of Statistics*, 40, 436-465.

[16] Bai, J., and K. Li (2016): "Maximum Likelihood Estimation and Inference for Approximate Factor Models of High Dimension," *Review of Economics and Statistics*, 98, 298-309.

[17] Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191-221.

[18] Bekker, P., and J. ten Berge (1997): "Generic Global Identification in Factor Analysis," *Linear Algebra and its Applications*, 264, 255-263.

[19] Bernal, P., N. Mittag, and J. Qureshi (2016): "Estimating Effects of School Quality Using Multiple Proxies," *Labour Economics*, 39, 1-10.

[20] Black D., and J. Smith (2006): "Estimating the Returns to College Quality with Multiple Proxies for Quality," *Journal of Labor Economics*, 24, 701-728.

[21] Bollen, K. (1989): *Structural Equations with Latent Variables*, John Wiley & Sons.

[22] Briggs, N., and R. MacCallum (2003): "Recovery of Weak Common Factors by Maximum Likelihood and Ordinary Least Squares Estimation," *Multivariate Behavioral Research*, 38, 25-56.

[23] Center for Human Resource Research (2006): *NLSY79 Child and Young Adult Data User's Guide.* Columbus, OH: Ohio State University.

[24] Chaudhuri, S., and E. Zivot (2011): "A New Method of Projection-Based Inference in GMM with Weakly Identified Nuisance Parameters," *Journal of Econometrics*, 164, 239-251.

[25] Chen, X., T. Christensen, K. O'Hara, and E. Tamer (2016): "MCMC Confidence Sets for Identified Sets," arXiv: 1605:00499v2.

[26] Chen, X., M. Ponomareva, and E. Tamer (2014): "Likelihood Inference in Some Finite Mixture Models," *Journal of Econometrics*, 182, 87-99.

[27] Cheng, X. (2015): "Robust Inference in Nonlinear Models with Mixed Identification Strength," *Journal of Econometrics*, 189, 207-228.

[28] Cho, J., and H. White (2007): "Testing for Regime Switching," *Econometrica 2007*, 75, 1671-1720.

[29] Cox, G. (2016a) "A Higher-Order Stochastic Expansion of M- and Z- Estimators," Unpublished Manuscript. [link]

[30] Cox, G. (2016b): "Generic Uniqueness of a Global Minimum," Unpublished Manuscript. [link]

[31] Cox, G. (2016c): "SM1: Limit Theorems for an Extremum Estimator in a Class of Generically Identified Models," Unpublished Manuscript. [link]

[32] Cox, G. (2016d): "SM2: Robust Inference for a Weak Factor," Unpublished Manuscript. [link]

[33] Cox, G. (2016e): "SM3: Robust Inference for Entangled Factors," Unpublished Manuscript. [link]

[34] Cunha, F., and J. Heckman (2008): "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Journal of Human Resources*, 43, 738-782.

[35] Cunha, F., J. Heckman, and S. Schennach (2010): "Estimating the Technology of cognitive and noncognitive skill formation," *Econometrica*, 78, 883-931.

[36] Dufour, J., and M. Taamouti (2005): "Projection-Based Statistical Inference in Linear Structural Models with Possibly Weak Instruments," *Econometrica*, 73, 1351-1365.

[37] Guggenberger, P., and R. Smith (2005): "Generalized Empirical Likelihood Estimators and Tests Under Partial, Weak, and Strong Identification," *Econometric Theory*, 21, 667-709.

[38] Han, S., and A. McCloskey (2016): "Estimation and Inference with a (Nearly) Singular Jacobian," Unpublished Manuscript.

[39] Heckman, J., R. Pinto, and P. Savelyev (2013): "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," *The American Economic Review*, 103, 2052-2086.

[40] Heywood, H. (1931): "On Finite Sequences of Real Numbers," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 134, 486-501.

[41] Kleibergen, F. (2005): "Testing Parameters in GMM Without Assuming that They Are Identified," *Econometrica*, 73, 1103-1123.

[42] Kleibergen, F. (2007): "Generalizing Weak Instrument Robust IV Statistics Towards Multiple Parameters, Unrestricted Covariance Matrices and Identification Statistics," *Journal of Econometrics*, 139, 181-216.

[43] Kleibergen, F. (2009): "Tests of Risk Premia in Linear Factor Models," *Journal of Econometrics*, 149, 149-173.

[44] Lawley, D., and A. Maxwell (1971): *Factor Analysis as a Statistical Method.* New York: American Elsevier Publishing Company, Inc.

[45] Ledermann, W. (1937): "On the Rank of the Reduced Correlation Matrix in Multiple-Factor Analysis," *Psychometrika*, 2, 85-93.

[46] Lekfuangfu, N. (2015): "Human Capital Investment: An Empirical Analysis of Incentives and Returns," Doctoral dissertation, UCL (University College London).

[47] McCloskey, A. (2012): "Bonferroni-Based Size-Correction for Nonstandard Testing Problems," Unpublished Manuscript.

[48] Onatski, A. (2012): "Asymptotics of the Principal Components Estimator of Large Factor Models with Weakly Influential Factors," *Journal of Econometrics*, 168, 244-258.

[49] Otsu, T. (2006): "Generalized Empirical Likelihood Inference for Nonlinear and Time Series Models Under Weak Identification," *Econometric Theory*, 22, 513-527.

[50] Pavan, R. (forthcoming): "On The Production of Skills and the Birth Order Effect," *Journal of Human Resources.*

[51] Rothenberg, T. (1971): "Identification in Parametric Models," *Econometrica*, 39, 577-591.

[52] Sargan, D. (1983): "Identification and Lack of Identification," *Econometrica*, 51,1605-1633.

[53] Shapiro, A. (1985): "Identifiability of Factor Analysis: Some Results and Open Problems," *Linear Algebra and its Applications*, 70, 1-7.

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

[55] Stock, J., and J. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68, 1055-1096.

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

[57] Ximénez, C. (2006): "A Monte Carlo Study of Recovery of Weak Factor Loadings in Confirmatory Factor Analysis," *Structural Equation Modeling*, 13, 587-614.

[58] Ximénez, C. (2007): "Effect of Variable and Subject Sampling on Recovery of Weak Factors in CFA," *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 3, 67-80.

[59] Ximénez, C. (2009): "Recovery of Weak Factor Loadings in Confirmatory Factor Analysis Under Conditions of Model Misspecification," *Behavior Research Methods*, 41, 1038-1052.

[60] Ximénez, C. (2015): "Recovery of Weak Factor Loadings When Adding the Mean Structure in Confirmatory Factor Analysis: A Simulation Study," *Frontiers in Psychology*, 6, Article 1943.

A Data Description

The variables used are a subset of the ones used by Cunha, Heckman, and Schennach (2010). The specification for children ages 0-2 years old uses: how often the child gets out of the house, the number of books the child has, how often mom reads to the child, the number of soft/role play toys, the number of push/pull toys, how often the child eats with mom/dad, and how often mom talks to the child from work. The specification for children ages 3-5 years old uses: how often the child gets out of the house, the number of books the child has, how often mom reads to the child, how often the child eats with mom/dad, number of magazines, whether the child has a tape recorder/CD player, and how often the child is taken to a museum. The specification for children ages 6-9 years old uses: the number of books the child has, how often mom reads to the child, how often the child eats with mom/dad, how often the child is taken to a museum, whether the family receives a daily newspaper, whether the child is taken to musical performances, and how often the child sees family friends. The specification for children ages 10-12 years old uses: the number of books the child has, how often the child eats with mom/dad, how often the child is taken to a

Category	0-2 Years	3-5 Years	10-12 Years
Cognitive Normalization:	Books	Books	Books
Noncognitive Normalization:	Outings	Outings	Complimented
Cognitive Investment 1:	Push/Pull Toys	Museum	Museum
Cognitive Investment 2:	Reads	Reads	Music

Table 5: Normalization variables and cognitive investment variables for ages 0-2, 3-5, and 10-12 years.

museum, whether the child is taken to musical performances, how often the child sees family friends, the number of times the child was praised last week, and the number of times the child was complimented last week.

Table 5 states which variables are used for the normalizations and cognitive investments for age specifications 0-2, 3-5, and 10-12 years-old.[46]

The data includes observations from 2810 firstborn white children born to female respondents in the NLSY dataset. These are the same observations used by Cunha, Heckman, and Schennach (2010), updated for newly available observations. More information on the dataset can be found in the description of the data given by Cunha, Heckman, and Schennach (2010), or in Center for Human Resource Research (2006).

B Example 2 with More Measurements

Robust Inference in Example 2 was specific to a model with five measurements. This appendix extends that analysis to more measurements. Let the model be:

$$X_i = \Lambda F_i + \epsilon_i,$$

[46] Books denotes the number of books the child has. Complimented denotes the number of times the child was complimented last week. Museum denotes how often the child is taken to a museum. Music denotes whether the child is taken to musical performances. Outings denotes how often the child gets out of the house. Push/Pull Toys denotes the number of push/pull toys. Reads denotes how often mom reads to the child.

where X_i is a p-vector with $p \geq 5$. F_i is still a 2-vector with covariance Σ and ϵ_i is a p-vector with covariance Φ, assumed to be diagonal. If we let

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \lambda_{11} & \lambda_{12} \\ \vdots & \vdots \\ \lambda_{p-2,1} & \lambda_{p-2,2} \end{bmatrix},$$

and let $\lambda_i = (\lambda_{i1}, \lambda_{i2})'$ for $i = 1, ..., p-2$, then we can write the covariance of X_i as:

$$\begin{bmatrix} \sigma_1^2 + \phi_1 & \sigma_{12} & \lambda_1' \Sigma e_1 & \lambda_2' \Sigma e_1 & \cdots & \lambda_{p-2}' \Sigma e_1 \\ {}'' & \sigma_2^2 + \phi_2 & \lambda_1' \Sigma e_2 & \lambda_2' \Sigma e_2 & \cdots & \lambda_{p-2}' \Sigma e_2 \\ {}'' & {}'' & \lambda_1' \Sigma \lambda_1 + \phi_3 & \lambda_1' \Sigma \lambda_2 & \cdots & \lambda_1' \Sigma \lambda_{p-2} \\ {}'' & {}'' & {}'' & \lambda_2' \Sigma \lambda_2' + \phi_4 & \cdots & \lambda_2' \Sigma \lambda_{p-2} \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ {}'' & {}'' & {}'' & {}'' & \cdots & \lambda_{p-2}' \Sigma \lambda_{p-2} + \phi_p \end{bmatrix}.$$

We can reparameterize this as:

$$\begin{bmatrix} \zeta_1 & \sigma_{12} & \beta_{11} & \beta_{21} & \beta_{31} & \beta_{41} & \cdots & \beta_{p-2,1} \\ {}'' & \zeta_2 & \beta_{12} & \beta_{22} & \beta_{32} & \beta_{42} & \cdots & \beta_{p-2,2} \\ {}'' & {}'' & \zeta_3 & \rho & h_{13}(\psi, \pi) + \rho \frac{\beta_{32}}{\beta_{22}} & h_{14}(\psi, \pi) + \rho \frac{\beta_{42}}{\beta_{22}} & \cdots & h_{1,p-2}(\psi, \pi) + \rho \frac{\beta_{p-2,2}}{\beta_{22}} \\ {}'' & {}'' & {}'' & \zeta_4 & h_{23}(\psi, \pi) + \rho \frac{\beta_{32}}{\beta_{12}} & h_{24}(\psi, \pi) + \rho \frac{\beta_{42}}{\beta_{12}} & \cdots & h_{2,p-2}(\psi, \pi) + \rho \frac{\beta_{p-2,2}}{\beta_{12}} \\ {}'' & {}'' & {}'' & {}'' & \zeta_5 & h_{34}(\psi, \pi) + \rho \frac{\beta_{32}\beta_{42}}{\beta_{12}\beta_{22}} & \cdots & h_{3,p-2}(\psi, \pi) + \rho \frac{\beta_{32}\beta_{p-2,2}}{\beta_{12}\beta_{22}} \\ {}'' & {}'' & {}'' & {}'' & {}'' & \zeta_6 & \cdots & h_{4,p-2}(\psi, \pi) + \rho \frac{\beta_{42}\beta_{p-2,2}}{\beta_{12}\beta_{22}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ {}'' & {}'' & {}'' & {}'' & {}'' & {}'' & \cdots & \zeta_p \end{bmatrix},$$

where

$$\begin{aligned} h_{1j}(\psi, \pi) &= \frac{\beta_{11}(\beta_{j1}\beta_{22} - \beta_{j2}\beta_{21})}{\pi\beta_{22} - \sigma_{12}\beta_{21}} + \rho \frac{\beta_{j2}\pi - \beta_{j1}\sigma_{12}}{\pi\beta_{22} - \sigma_{12}\beta_{21}} - \rho \frac{\beta_{j2}}{\beta_{22}} \\ h_{2j}(\psi, \pi) &= \frac{\beta_{21}(\beta_{j1}\beta_{12} - \beta_{j2}\beta_{11})}{\pi\beta_{12} - \sigma_{12}\beta_{11}} + \rho \frac{\beta_{j2}\pi - \beta_{j1}\sigma_{12}}{\pi\beta_{12} - \sigma_{12}\beta_{11}} - \rho \frac{\beta_{j2}}{\beta_{12}} \\ h_{ij}(\psi, \pi) &= \frac{\beta_{11}(\sigma_{12}\beta_{i1} - \beta_{i2}\pi)(\beta_{21}\beta_{j2} - \beta_{11}\beta_{j1})}{(\pi\beta_{12} - \sigma_{12}\beta_{11})(\pi\beta_{22} - \sigma_{12}\beta_{21})} + \frac{\beta_{21}(\sigma_{12}\beta_{j1} - \beta_{j2}\pi)(\beta_{11}\beta_{i2} - \beta_{12}\beta_{i1})}{(\pi\beta_{12} - \sigma_{12}\beta_{11})(\pi\beta_{22} - \sigma_{12}\beta_{21})} \end{aligned}$$

$$+\rho\frac{(\pi\beta_{i2}-\sigma_{12}\beta_{i1})(\pi\beta_{j2}-\sigma_{12}\beta_{j1})}{(\pi\beta_{12}-\sigma_{12}\beta_{11})(\pi\beta_{22}-\sigma_{12}\beta_{21})}-\rho\frac{\beta_{i2}\beta_{j2}}{\beta_{12}\beta_{22}},$$

for $j = 3, ..., p-2$ and $i = 3, ..., j-1$.

Then $h(\psi, \pi)$ does not depend on π when

$$0 = \begin{pmatrix} \beta_{11}\beta_{22} - \beta_{12}\beta_{21} \\ \beta_{11}\beta_{32} - \beta_{12}\beta_{31} \\ \vdots \\ \beta_{11}\beta_{j2} - \beta_{12}\beta_{j1} \\ \vdots \\ \beta_{11}\beta_{p-2,2} - \beta_{12}\beta_{p-2,1} \end{pmatrix} = \gamma(\beta).$$

Thus, we can define t_n to be:

$$t_n = \left(\frac{n}{p-3}\gamma(\hat{\beta}_n)'(B(\hat{\beta}_n)\hat{\text{Var}}_{\beta\beta}(\hat{\theta}_n)B(\hat{\beta}_n)')^{-1}\gamma(\hat{\beta}_n)\right)^{1/2},$$

where $B(\beta) = \frac{\partial}{\partial\beta}\gamma(\beta)$ and $\hat{\text{Var}}_{\beta\beta}(\hat{\theta})$ is an estimator of the upper left block of the asymptotic variance of $\hat{\theta}$. Then t_n is a test statistic for testing $H_0 : \gamma(\beta) = 0$. We can calculate robust critical values using straightforward extensions to the formulas for H, Y_1, $d(\pi)$, and $d_1(\pi)$ plugged into the formulas for $Z(\pi)$ and $\xi(\pi)$.

This extends the analysis of Example 2 to allow for an arbitrary number of measurements.

C More Factors: Identification Reduction

This section gives suggestions for robust inference in factor models with more than two factors.

As the number of factors increases, the rank condition becomes more complicated. In fact, Assumption Rank Condition is sufficient but not necessary when the number of factors is three or larger. This means that applying the robust inference method of Section 2 directly requires: (1) specifying a rank condition that is both necessary and sufficient for identification, (2) reparameterizing the model to fit Definition Parameter Spaces and Definition Objective Functions, and (3) characterizing the asymptotic distribution of a test statistic along sequences of parameters converging to points of rank condition failure.

I have two suggestions that may simplify this process. The first is to notice that a complicated rank condition often has multiple ways that it can fail, and not all of them may be relevant for a particular application. By imposing appropriate assumptions, the rank condition can be significantly simplified. Example 2 already takes advantage of this. Example 2 assumes that at least two of the non-normalized measurements have nonzero factor loadings for the first factor, which eliminates the possibility of two weak factors.[47] This reduces the sources of rank condition failure to (a) a weak fifth measurement, (b) a weak second factor, or (c) two entangled factors. This is appropriate in the parental investments application because it is reasonable to assume that parental investment in cognitive skills is a strong factor. For that application, the relevant way that the rank condition can fail is due to entangled factors. Thus, additional assumptions that are appropriate for the application can simplify a complicated rank condition.

The second suggestion is to notice that specification changes in the model can change the rank condition. In particular, imposing block diagonality on the factor loadings may reduce the rank condition of the full model to rank conditions on each block. Suppose the model is given by:

$$X_i = \Lambda F_i + \epsilon_i.$$

Also suppose that Λ is block diagonal, so that

$$\Lambda = \begin{bmatrix} I_{m_1} & 0 \\ \Lambda_1 & 0 \\ 0 & I_{m_2} \\ 0 & \Lambda_2 \end{bmatrix},$$

where the number of factors in each block is m_j and the number of measurements in each block is p_j for $j = 1, 2$. This specification has only two blocks, but can be easily extended to more blocks. Furthermore, assume that the variance of the errors satisfies:

$$\Phi = \begin{bmatrix} \text{diag}(\phi_1, ..., \phi_{m_1}) & 0 & 0 & \Phi_{14} \\ 0 & \text{diag}(\phi_{m_1+1}, ..., \phi_{p_1}) & \Phi_{23} & \Phi_{24} \\ 0 & \Phi_{32} & \text{diag}(\phi_{p_1+1}, ..., \phi_{p_1+m_2}) & 0 \\ \Phi_{41} & \Phi_{42} & 0 & \text{diag}(\phi_{p_1+m_2+1}, ..., \phi_p) \end{bmatrix},$$

where $\text{diag}(a_1, ..., a_m)$ denotes a $m \times m$ diagonal matrix with diagonal entries given by $a_1, ..., a_m$. $\Phi_{41} = \Phi_{14}'$ denotes a $p_2 - m_2 \times m_1$ matrix of parameters, $\Phi_{32} = \Phi_{23}'$ denotes a $m_2 \times p_1 - m_1$ matrix of parameters, and $\Phi_{42} = \Phi_{24}'$ denotes a $p_2 - m_2 \times p_1 - m_1$ matrix of parameters. Also let $\Phi_{11} = \text{diag}(\phi_1, ..., \phi_{m_1})$, $\Phi_{22} = \text{diag}(\phi_{m_1+1}, ..., \phi_{p_1})$, $\Phi_{33} = \text{diag}(\phi_{p_1+1}, ..., \phi_{p_1+m_2})$, and $\Phi_{44} = \text{diag}(\phi_{p_1+m_2+1}, ..., \phi_p)$. Allowing for these off-diagonal terms in the covariance matrix of the errors has two purposes. The first is to relax the assumption of the diagonal error covariance matrix, and the second is to reduce the rank

[47] See the definition of the parameter space in Supplemental Materials 3.

condition. Also let the variance of the factors be denoted by

$$\Sigma = \left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array} \right].$$

In this model, the covariance matrix of the observables can be written as

$$\begin{aligned} Cov(X_i) &= \Lambda\Sigma\Lambda' + \Phi \\ &= \left[\begin{array}{cccc} \Sigma_{11} + \Phi_{11} & \Sigma_{11}\Lambda_1' & \Sigma_{12} & \Sigma_{12}\Lambda_2' + \Phi_{14} \\ \Lambda_1\Sigma_{11} & \Lambda_1\Sigma_{11}\Lambda_1' + \Phi_{22} & \Lambda_1\Sigma_{12} + \Phi_{23} & \Lambda_1\Sigma_{12}\Lambda_2' + \Phi_{24} \\ \Sigma_{21} & \Sigma_{21}\Lambda_1' + \Phi_{32} & \Sigma_{22} + \Phi_{33} & \Sigma_{22}\Lambda_2' \\ \Lambda_2\Sigma_{21} + \Phi_{41} & \Lambda_2\Sigma_{21}\Lambda_1' + \Phi_{42} & \Lambda_2\Sigma_{22} & \Lambda_2\Sigma_{22}\Lambda_2' + \Phi_{44} \end{array} \right]. \end{aligned}$$

This equation can be solved for the parameters in $\Sigma_{11}, \Sigma_{12}, \Sigma_{22}, \Lambda_1, \Lambda_2, \Phi_{11}, \Phi_{22}, \Phi_{33}, \Phi_{44}, \Phi_{14}$, Φ_{23}, and Φ_{24} if and only if we can solve for the parameters in the two smaller factor models, given by

$$\begin{aligned} X_{1i} &= \Lambda_1 F_{1i} + \epsilon_{1i} \\ X_{2i} &= \Lambda_2 F_{2i} + \epsilon_{2i}, \end{aligned}$$

where X_{1i} denotes the first p_1 variables in X_i, X_{2i} denotes the last p_2 variables in X_i, F_{1i} denotes the first m_1 factors in F_i, F_{2i} denotes the last m_2 factors in F_i, ϵ_{1i} denotes the first p_1 errors in ϵ_i, and ϵ_{2i} denotes the last p_2 errors in ϵ_i. That is, if and only if we can solve for the parameters in

$$\begin{aligned} Cov(X_{1i}) &= \left[\begin{array}{cc} \Sigma_{11} + \Phi_{11} & \Sigma_{11}\Lambda_1' \\ \Lambda_1\Sigma_{11} & \Lambda_1'\Sigma_{11}\Lambda_1 + \Phi_{22} \end{array} \right] \\ Cov(X_{2i}) &= \left[\begin{array}{cc} \Sigma_{22} + \Phi_{33} & \Sigma_{22}\Lambda_2' \\ \Lambda_2\Sigma_{22} & \Lambda_2\Sigma_{22}\Lambda_2' + \Phi_{44} \end{array} \right]. \end{aligned}$$

These two equations are the covariance matrices for the smaller factor models, and can be solved if and only if rank conditions hold on Λ_1 and Λ_2. Thus, identification in the larger factor model is determined by rank conditions on Λ_1 and Λ_2. Robust inference with respect to the rank conditions on Λ_1 and Λ_2 can then be done using the sequential peeling argument of Cheng (2015), which extends to an arbitrary number of blocks.

This shows that under block diagonality of the factor loadings and allowing for some nonzero off-diagonal terms in the error covariance matrix, the identification status of a large factor model can be reduced to the identification status of smaller factor models.

D Notation for Tensors

The proof of Theorem 2 requires a higher-order expansion of the objective function. The higher-order terms involve higher-order derivatives that are stored in higher-order matrices called tensors. Tensors can be manipulated just like matrices, but with some additional notation. For example, if A is a $d_\psi \times d_\psi \times d_\psi$ tensor, and $i_1, i_2, i_3 \in \{1, ..., d_\psi\}$, then A_{i_1,i_2,i_3} denotes the $(i_1, i_2, i_3)^{\text{th}}$ element of A. For example, if $f(x)$ is a three times continuously differentiable function, then we can define a three tensor of third order partial derivatives to be $A_{i_1,i_2,i_3} = \frac{\partial^3}{\partial x_{i_1} \partial x_{i_2} \partial x_{i_3}} f(x)$ for all $i_1, i_2, i_3 \in \{1, ..., d_x\}$. This example generalizes to higher derivatives and higher dimensions of the tensor.

The space of tensors contains a natural inner product, corresponding to the Frobenius norm for matrices. If A and B are d_ψ^m tensors, then the inner product between A and B is defined to be

$$\langle A; B \rangle = \sum_{i_1=1}^{d_\psi} \cdots \sum_{i_m=1}^{d_\psi} A_{i_1,...,i_m} B_{i_1,...,i_m}.$$

This inner product also defines a norm, $||A||$, that satisfies the Cauchy-Schwarz inequality. This norm also satisfies $||A \otimes B|| = ||A|| ||B||$. This bracket notation can also be defined for A and B with different dimensions. For example, if A is a $d_\psi^{m_A}$ tensor and B is a $d_\psi^{m_B}$ tensor, with $m_A > m_B$, then $\langle A; B \rangle$ is defined to be a $d_\psi^{m_A - m_B}$ tensor with $(i_1, ..., i_{m_A - m_B})^{\text{th}}$ element equal to:

$$\sum_{i_{m_A - m_B + 1}=1}^{d_\psi} \cdots \sum_{i_{m_A}=1}^{d_\psi} A_{i_1,...,i_{m_A - m_B}, i_{m_A - m_B + 1},...,i_{m_A}} B_{i_{m_A - m_B + 1},...,i_{m_A}}.$$

The convention is that the product is taken over the last m_B dimensions of A. This generalizes multiplication of a matrix by a vector. An important consequence of these definitions is the fact that the norm is compatible over this multiplication, meaning that for any tensors A, B with the dimension of A greater than or equal to the dimension of B, $||\langle A; B \rangle|| \leq ||A|| ||B||$.

Another important operation is tensor multiplication. If A is a $d_\psi^{m_A}$ tensor and B is a $d_\psi^{m_B}$ tensor, then $A \otimes B$ is a $d_\psi^{m_A + m_B}$ tensor that is indexed by the dimensions of A first and the dimensions of B second. This can be generalized to allow for more products, $\otimes_j A_j$, and powers, $A^{\otimes 2}$, in the natural way.

The final operation is a generalization of the transpose. If A is a d_ψ^m tensor, let $\circlearrowleft(A)$ denote the tensor with $(i_1, ..., i_m)^{\text{th}}$ element given by $A_{i_2,...,i_m,i_1}$. This operation moves the last dimension of A to be first and all of the other dimensions are shifted back one. This operation can be repeated to cycle through the dimensions of A.

Finally, these operations can be defined for tensors that have different lengths for different dimensions. In this case, one additional extension occurs with the bracket notation. We can combine the bracket with the tensor product in the following way. If A is a $d_\psi^{m_1} \times d_h^{m_2}$ tensor and B is a $d_\psi^{m_3} \times d_h^{m_4}$ tensor and $m_2 \geq m_4$, then let $\langle A; B\rangle_{d_h}$ denote the $d_\psi^{m_1+m_3} \times d_h^{m_2-m_4}$ tensor that takes the product over the final m_4 dimensions that are length d_h, while the remaining dimensions are stacked, as in the tensor product.