Maximum Likelihood Inference in Weakly Identified DSGE Models.

By Isaiah Andrews[1] and Anna Mikusheva [2] [3]

April 14, 2011

Abstract

This paper examines the issue of weak identification in maximum likelihood, motivated by problems with estimation and inference in a multi-dimensional, non-linear DSGE model. We suggest a test for a simple hypothesis concerning the full parameter vector which is robust to weak identification. We also suggest a test for a composite hypothesis regarding a sub-vector of parameters. The suggested test is shown to be asymptotically exact when the nuisance parameter is strongly identified, and in some cases when the nuisance parameter is weakly identified. We pay particular attention to the question of how to estimate Fisher's information, and make extensive use of martingale theory.

Key words: weak identification, maximum likelihood, score test, $C(\alpha)-$ test

1 Introduction

Recent years have witnessed the rapid growth of the empirical literature on the highly parameterized micro-founded macro models known as Dynamic Stochastic General Equilibrium (DSGE) models. A number of papers in this literature have considered estimating these models by maximum likelihood (see for example Ingram, Kocherlakota and Savin (1994), Ireland (2004), Lindé (2005), and McGrattan, Rogerson and Wright (1997)). More recently, Bayesian estimation has become increasingly popular, due in large part to the difficulty of maximum likelihood estimation in many DSGE models. As Fernández-Villaverde (2010) points out in his survey of DSGE estimation, "likelihoods of DSGE models are full of local maxima and minima and of nearly flat surfaces... the standard errors of the estimates are notoriously difficult to compute and their asymptotic

[1]Department of Economics, M.I.T., 50 Memorial Drive, Building E52, Cambridge, MA, 02142. Email: iandrews@mit.edu. Financial support from the Ford Foundation is gratefully acknowledged.

[2]Department of Economics, M.I.T., 50 Memorial Drive, Building E52, Cambridge, MA, 02142. Email: amikushe@mit.edu. Financial support from the Castle-Krob Career Development Chair is gratefully acknowledged.

[3]We would like to thank Whitney Newey, Jim Stock and seminar participants at Columbia, Harvard-MIT, Rice and Texas A& M, for helpful comments. We would also like to thank Frank Kleibergen and Sophocles Mavroeidis for informative discussion of their approach in GMM.

distribution a poor approximation to the small sample one." The poor performance of maximum likelihood estimation has fueled growing concerns about poor identification in many DSGE models (see Canova and Sala (2009), Guerron-Quintana, Inoue and Kilian (2009), and Iskrev (2010)).

In this paper, we consider the problem of weak identification in dynamic models estimated by maximum likelihood. Weak identification arises when the amount of information in the data about some parameter or group of parameters is small and is generally modeled in such a way that information about parameters accumulates slowly along some dimensions. This leads to the breakdown of the usual asymptotics for maximum likelihood, with the asymptotic distributions for the maximum likelihood estimator and the standard LR, LM, and Wald statistics providing a poor approximation to their finite sample behavior. This is distinct from loss of point identification. We assume throughout that the models we consider are point identified, and thus that changing the value of any parameter changes the distribution of the data, though the effect will be small for some parameters.

We focus on the problem of testing and confidence set construction in this context. In our view there are two main approaches to inference in models where identification may be weak. One is to create a two-step procedure, where one first differentiates (via a pre-test) between weakly and strongly identified models and then chooses a procedure based on the test result. We take the other approach. Rather than looking for a test for weak identification as such, we instead attempt to construct a test for parameters which is robust to weak identification. The ideal procedure should satisfy two conditions. First, it should control size well if identification is weak, and second, it should be asymptotically equivalent to the classical MLE tests if identification is strong. If such a procedure exists, it renders pretests unnecessary and, in general, inferior given the size problems endemic to multiple testing procedures.

We view this approach as analogous to the modern treatment of testing in the presence of potential heteroscedasticity. While in the past it was common to use pretests for heteroscedasticity, current empirical practice is to simply use standard errors (such as those of White (1980)) which are correct asymptotically regardless of whether or not the data is heteroscedastic. Likewise, in weak instrumental variables regression (weak IV) there are tests available which have correct asymptotic size under weak identification and (at least for the case of one endogenous variable) at least as much power as the

classical procedures under strong identification. Unlike the case of heteroscedasticity, where weighted least squares could potentially improve precision, in weak IV the outcome of a pretest cannot be used to increase power, so there is even less reason to use a pretest-based procedure.

We construct a robust test in two steps. First, we suggest a test for a simple hypothesis on the full parameter vector. This test is robust to weak identification and is asymptotically equivalent to the classical Lagrange Multiplier (LM) test when identification is strong. The assumptions needed for this result are extremely weak and cover a large number of cases, including weak IV, an ARMA(1,1) with nearly canceling roots, a weakly identified binary choice model and weakly identified exponential family models, for example VARs with weakly identified structural parameters. The proof for this test makes extensive use of martingale theory, particularly the fact that the score (i.e. the gradient of the log likelihood) is a martingale when evaluated at the true parameter value.

Next, we turn to the problem of testing a subset of parameters without restricting the remaining parameters. Creation of such tests is critical for the construction of confidence sets, given that the common practice in applied work is to report a separate confidence interval for each element of the parameter vector. Constructing a test satisfying our first requirement, that is, one that controls size well under weak identification, is straightforward using our test for the full parameter vector and the projection method. However, simultaneously satisfying the second condition, asymptotic equivalence to classical tests under strong identification, is a much more challenging problem which (to the best of our knowledge) has not been fully solved even for many simpler models.

The test which we suggest for a subset of parameters is asymptotically equivalent to Neyman's $C(\alpha)$ test when identification is strong. We show that the suggested test has a χ^2 asymptotic distribution so long as the nuisance parameter (i.e. the part of the parameter vector which we are not testing) is strongly identified, without any assumption about the strength of identification of the tested parameter. We also show that the suggested test has the correct asymptotic size in some cases where the nuisance parameter is weakly identified. In particular we consider the case of an exponential family model where part of the nuisance parameter is weakly identified and enters linearly while no assumption is made on the strength of identification of the tested parameter. As a special case we examine weak IV with one endogenous variable when the nuisance parameter is weakly identified.

In addition to these theoretical results, we report simulation results showing that our proposed test maintains size well in a simple nonlinear model and an ARMA(1,1) model with nearly canceling roots. We also show the applicability of our results to a basic DSGE model.

Relation to the Literature on Weak Identification The literature on weak identification is quite large. The most-studied and best-understood case is that of weak instrumental variables estimation. For a comprehensive survey of the literature on this topic, see Stock, Wright, and Yogo (2002). The weak identification framework was generalized to GMM by Stock and Wright (2000), who represented weak identification using an asymptotic embedding in which the objective function becomes flat along some dimensions as the sample grows. While we make use of a similar embedding to demonstrate the applicability of our assumptions in an exponential family model, it is in no way necessary for our results, and we remain quite agnostic about the process generating the data. An alternative embedding for weak identification is introduced in Andrews and Cheng (2009).

Making use of their embedding, Stock and Wright (2000) introduce tests for GMM which are robust to weak identification. They consider two types of test: a test for the full parameter vector (i.e. for a simple hypothesis) and a test for a sub-parameter for the case where the nuisance parameter is well identified. Kleibergen and Mavroeidis (2009) suggest adaptations of the Stock and Wright (2000) S and Kleibergen (2005) KLM tests for a sub-parameter for the case when the nuisance parameter is weakly identified, which yield conservative tests asymptotically. While the statistics we consider are in many ways similar to those considered by Stock and Wright (2000), Kleibergen (2005), and Kleibergen and Mavroeidis (2009), their results do not in general apply to the context we consider as the variance of the moment condition (the score of the log likelihood) becomes degenerate asymptotically, violating one of their assumptions.

The issue of weak identification in DSGE models was first introduced by Canova and Sala (2009), who point out that the objective functions implied by many DSGE models are nearly flat in some directions. A weak identification-robust inference procedure for DSGE models based on likelihood analysis was introduced by Guerron-Quintana Inoue and Killian (2009). Their approach makes extensive use of the projection method for constructing confidence sets for the structural parameters which, given the high dimension

of the parameter space in many DSGE models, has the potential to introduce a substantial amount of conservativeness in many applications. Dufour, Khalaf and Kichian (2009) offer another approach for Full-Information analysis of weakly identified DSGE models, based on the Anderson-Rubin statistic. Another paper on weak identification in DSGE models is Iskrev (2010), which attempts to asses the quality of identification in DSGE models by considering the degeneracy of the Hessian of the log likelihood. There are also a few papers discussing point-identification in DSGE models, which are unrelated to our paper as we assume point-identification. We refer the interested reader to Komunjer and Ng (2009) for an example of this literature.

Relation to the Classical MLE Literature The other major literature to which our paper is connected is the classical Statistics literature on maximum likelihood. This classical literature began in the i.i.d. context and was generalized considerably by Le Cam (see Le Cam and Yang (2000)), allowing the use of MLE in a wide array of problems, including those with dependent data. The application of ML to dependent data was further explored by a number of other authors, including Silvey (1961), Crowder (1976), Heijmans and Magnus (1986) and Jeganathan (1995). Our approach is particularly informed by the strand of this literature which focuses on the martingale properties of the log likelihood and their implications for the asymptotics of the MLE, and especially by Bhat (1974) and Hall and Heyde (1980).

The weakly identified dynamic models we consider differ from those in this classical literature in that the normalized second derivative of the log likelihood may not converge to a constant (or, if normalized to converge to a constant, may be singular asymptotically). As a result, these models fall outside of the classes considered by the previous literature (to take a non-dynamic example, it can be shown that the standard weak IV model is not Locally Asymptotically Quadratic, and thus is not subject to the results of Le Cam). Some additional complications in the DSGE context include the fact that the parameter space is in general quite large and that analytic expressions for the log likelihood are in general unavailable, though the likelihood can be evaluated numerically.

Structure of the paper Section 2 introduces our notation as well as some results from martingale theory; it also discusses the difference between two alternative measures of information and illustrates this difference in several examples. Section 3 suggests a

test for the full parameter vector. Section 4 discusses the problem of testing a composite hypothesis about a sub-parameter, and introduces a statistic for such a test. Section 5 proves that our sub-vector test is valid when the nuisance parameter is strongly identified without any assumption on the strength of identification of the tested parameter. Section 6 shows that this result can be extended to some cases in which the nuisance parameter is weakly identified. Simulations supporting our theoretical results are provided in Section 7.

Proofs of secondary importance and demonstrations that the assumptions of the paper hold in our examples are placed in the Supplementary Appendix, which can be found on Anna Mikusheva's website.[4] In particular, the proofs of the statement from Section 5.3 and Lemma 3 appear in the Supplementary Appendix.

Throughout the rest of the paper, Id_k is the $k \times k$ identity matrix, $\mathbb{I}\{\cdot\}$ is the indicator-function, $[\cdot]$ stands for the quadratic variation of a martingale and $[\cdot,\cdot]$ for the joint quadratic variation of two martingales, $\Rightarrow$ denotes weak convergence (convergence in distribution), while $\to^p$ stands for convergence in probability.

2 Martingale Methods in Maximum Likelihood

Let X_T be the data available at time T. In general, we assume that $X_T = (x_1, ..., x_T)$. Let $\mathcal{F}_t$ be a sigma-algebra generated by $X_t = (x_1, ..., x_t)$. We assume that the log likelihood of the model,

$$\ell(X_T;\theta) = \log f(X_T;\theta) = \sum_{t=1}^{T} \log f(x_t|\mathcal{F}_{t-1};\theta),$$

is known up to the k-dimensional parameter θ, which has true value θ_0. We further assume that $\ell(X_T;\theta)$ is twice continuously differentiable with respect to θ, and that the class of likelihood gradients $\left\{\frac{\partial}{\partial\theta'}\ell(X_T;\theta) : \theta \in \Theta\right\}$ and the class of second derivatives $\left\{\frac{\partial^2}{\partial\theta\partial\theta'}\ell(X_T;\theta)\right\}$ are both locally dominated integrable.

Our main object of study will be the score function,

$$S_T(\theta) = \frac{\partial}{\partial\theta'}\ell(X_T,\theta) = \sum_{t=1}^{T} \frac{\partial}{\partial\theta'} \log f(x_t|\mathcal{F}_{t-1};\theta),$$

where $s_t(\theta) = S_t(\theta) - S_{t-1}(\theta) = \frac{\partial}{\partial\theta'} \log f(x_t|\mathcal{F}_{t-1};\theta)$ is the increment of the score. Under

[4] https://econ-www.mit.edu/files/6648

the assumption that we have correctly specified the model, the expectation of $s_t(\theta_0)$ conditional on all information up to $t-1$ is equal to zero,

$$E\left(s_t(\theta_0)|\mathcal{F}_{t-1}\right) = 0 \quad a.s. \tag{1}$$

This in turn implies that the score taken at the true parameter value, $S_t(\theta_0)$, is a martingale with respect to filtration $\mathcal{F}_t$. One way to view (1) is as a generalization of the first informational equality, which in i.i.d. models states that $E\left[s_t(\theta_0)\right] = 0$, to the dynamic context. To derive this equality, note that $s_t(\theta_0) = \frac{1}{f(x_t|\mathcal{F}_{t-1};\theta_0)} \frac{\partial}{\partial \theta'} f(x_t|\mathcal{F}_{t-1};\theta_0)$,

$$E(s_t(\theta_0)|\mathcal{F}_{t-1}) = \int s_t(\theta_0) f(x_t|\mathcal{F}_{t-1};\theta_0) dx_t = \int \frac{\partial}{\partial \theta'} f(x_t|\mathcal{F}_{t-1};\theta_0) dx_t = 0.$$

This observation is due to Silvey (1961).

Similarly, the second informational equality also generalizes to the dependent case. In the i.i.d. case, this equality states that we can calculate Fisher's information using either the Hessian of the log likelihood or the outer product of the score, i.e.

$$\mathcal{I}(\theta_0) = -E\left(\frac{\partial^2}{\partial\theta\partial\theta'} \log f(x_t;\theta_0)\right) = E\left(\frac{\partial}{\partial\theta'} \log f(x_t;\theta_0) \frac{\partial}{\partial\theta} \log f(x_t;\theta_0)\right). \tag{2}$$

Fisher's information plays a key role in the classical asymptotics for maximum likelihood, as it is directly related to the asymptotic variance of the MLE, and (2) suggests two different ways of estimating it which are asymptotically equivalent in the classical context. To generalize (2) to the dynamic context, following Barndorff-Nielsen and Sorensen (1991), we introduce two measures of information based on observed quantities:

- *Observed information*: the negative Hessian of the log-likelihood,
$$I_T(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta'} \ell(X_T;\theta) = \sum_{t=1}^{T} i_t(\theta),$$
where $i_t(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta'} \log f(x_t|X_{t-1};\theta)$;

- *Incremental observed information:* the quadratic variation of the score of the log likelihood,
$$J_T(\theta) = [S(\theta)]_T = \sum_{t=1}^{T} s_t(\theta) s_t'(\theta),$$
where as before $s_t(\theta)$ is the increment of $S_T(\theta)$.

Using these definitions, let $A_T(\theta) = J_T(\theta) - I_T(\theta)$ be the difference between the two measures of observed information. The second informational equality implies that $A_t(\theta_0)$ is a martingale with respect to $\mathcal{F}_t$. Specifically, the increment of $A_t(\theta_0)$ is $a_t(\theta_0) = A_t(\theta_0) - A_{t-1}(\theta_0)$,

$$a_t(\theta_0) = \frac{\partial^2}{\partial\theta\partial\theta'} \log f(x_t|X_{t-1};\theta_0) + \frac{\partial}{\partial\theta'} \log f(x_t|X_{t-1};\theta_0) \frac{\partial}{\partial\theta} \log f(x_t|X_{t-1};\theta_0),$$

and an argument similar to that for the first informational equality gives us that $E(a_t|\mathcal{F}_{t-1}) = 0$ a.s.

In the classical context, $I_T(\theta_0)$ and $J_T(\theta_0)$ are asymptotically equivalent, which plays a key role in the asymptotics of maximum likelihood. In the i.i.d. case, for example, the law of large numbers implies that $\frac{1}{T} I_T(\theta_0) \to^p -E\left(\frac{\partial^2}{\partial\theta\partial\theta'} \log f(x_t,\theta_0)\right) = \mathcal{I}(\theta_0)$ and $\frac{1}{T} J_T(\theta_0) \to^p E\left(\frac{\partial}{\partial\theta'} \log f(x_t,\theta_0) \frac{\partial}{\partial\theta} \log f(x_t,\theta_0)\right) = \mathcal{I}(\theta_0)$. As a result of this asymptotic equivalence, the classical literature in the i.i.d. context uses these two measures of information more or less interchangeably.

The classical literature in the dependent context makes use of a similar set of conditions to derive the asymptotic properties of the MLE, focusing in particular on the asymptotic negligibility of $A_T(\theta_0)$ relative to $J_T(\theta_0)$. For example, Hall and Heyde (1980) show that for θ scalar, if higher order derivatives of the log-likelihood are asymptotically unimportant, $J_T(\theta_0) \to \infty$ a.s., and $\limsup_{T\to\infty} J_T(\theta_0)^{-1}|A_T(\theta_0)| < 1$ *a.s.*, then the MLE for θ is strongly consistent. If moreover, $J_T(\theta_0)^{-1} I_T(\theta_0) \to 1$ *a.s.*, then the ML estimator is asymptotically normal and $J_T(\theta_0)^{\frac{1}{2}}(\hat{\theta} - \theta_0) \Rightarrow N(0,1)$.

We depart from this classical approach in that we consider weak identification. Weak identification arises when information is small along some dimension, which we model by using an embedding such that Fisher's information is degenerate asymptotically. Similar embeddings have been used to study weak identification in other contexts, including the Weak Instrument asymptotics introduced by Staiger and Stock (1997), and the Weak GMM asymptotics of Stock and Wright (2000). In such an embedding the difference between our two measures of information is important, and $A_T(\theta_0)$ is no longer negligible asymptotically compared to observed incremental information $J_T(\theta_0)$, as demonstrated in the weak IV example below.

Example 1 We assume a reduced form model with normal errors:

$$\left\{ \begin{array}{c} y_t = \beta\pi' z_t + u_t \\ x_t = \pi' z_t + v_t \end{array} \right. , \left(\begin{array}{c} u_t \\ v_t \end{array} \right) \sim i.i.d.\ N(0, Id_2),$$

We take z_t to be a $k-$dimensional set of instruments, while β is the parameter of interest and π is a $k \times 1$ vector of nuisance parameters. Our assumption that the errors have known covariance matrix equal to Id_2 is not restrictive, since u_t and v_t are reduced form (rather than structural) errors, and thus are well-estimable. The analysis is done conditional on the instruments z_t, and for simplicity we assume that the data generating process for z_t is such that it satisfies a law of large numbers. Following the approach laid out by Staiger and Stock (1997), we represent weak identification by modeling π as local to zero, that is $\pi = \frac{1}{\sqrt{T}}C$, so π is drifting to zero as the sample grows.

Let $Y = (y_1, ..., y_T)'$, $X = (x_1, ..., x_T)'$ be $T \times 1$ and $Z = (z_1, ..., z_T)'$ be $T \times k$. In this model, we have the following log-likelihood:

$$\ell_T(\beta, \pi) = const - \frac{1}{2}(Y - \beta Z\pi)'(Y - \beta Z\pi) - \frac{1}{2}(X - Z\pi)'(X - Z\pi).$$

The score is

$$S_\beta(\theta) = \pi' Z'(Y - \beta Z\pi);\ S_\pi(\theta) = \beta Z'(Y - \beta Z\pi) + Z'(X - Z\pi).$$

Finally, the two measures of information are:

$$I_T(\theta_0) = -\frac{\partial^2}{\partial\theta\partial\theta'}\ell_T = \left(\begin{array}{cc} \pi' Z'Z\pi & \beta\pi' Z'Z - U'Z \\ \beta Z'Z\pi - Z'U & (1+\beta^2)Z'Z \end{array} \right);$$

$$J_T(\theta_0) = [S]_T = \left(\begin{array}{cc} \pi' \sum_t u_t^2 z_t z_t' \pi & \pi' \sum_t u_t(\beta u_t + v_t) z_t z_t' \\ \sum_t u_t(\beta u_t + v_t) z_t z_t' \pi & \sum_t (\beta u_t + v_t)^2 z_t z_t' \end{array} \right).$$

Under the weak instrument embedding $\pi = \frac{1}{\sqrt{T}}C$, we can use normalizing matrix $K_T = diag(1, \frac{1}{\sqrt{T}}, \ldots, \frac{1}{\sqrt{T}})$ to get a non-trivial limit for both information matrices:

$$K_T J_T(\theta_0) K_T \to^p \left(\begin{array}{cc} C'Q_Z C & \beta C'Q_Z \\ \beta Q_Z C & (1+\beta^2)Q_Z \end{array} \right);$$

$$K_T I_T(\theta_0) K_T \Rightarrow \left(\begin{array}{cc} C'Q_Z C & \beta C'Q_Z - \xi' \\ \beta Q_Z C - \xi & (1+\beta^2)Q_Z \end{array} \right).$$

To derive these expressions we have used a law of large numbers, $\frac{1}{T}Z'Z \to^p Q_Z$, and

a central limit theorem, $\frac{1}{\sqrt{T}}Z'U \Rightarrow \xi = N(0, Q_Z)$. Notice that, under weak instrument asymptotics, there is a difference between the two information matrices (i.e. the addition of the term $-\xi$ to the off-diagonal elements of $K_T J_T(\theta_0) K_T$), whereas for the strong IV case ($\pi \neq 0$ and fixed) we have that $J_T^{-1} I_T \to^p Id_2$. □

Example 2 Another well-known example of weak identification is the ARMA model with nearly canceling roots. Below we use the formulation of this model from Andrews and Cheng (2009). The relevance of this model to DSGE estimation is discussed in Schorfheide (2010).

$$Y_t = (\pi + \beta) Y_{t-1} + e_t - \pi e_{t-1}, \quad e_t \sim i.i.d. N(0,1).$$

The true value of parameter $\theta_0 = (\beta_0, \pi_0)'$ satisfies the following restrictions $|\pi_0| < 1$, $\beta_0 \neq 0$ and $|\pi_0 + \beta_0| < 1$, which guarantee that the process is stationary and invertible. For simplicity we assume that $Y_0 = 0$ and $e_0 = 0$, though due to stationarity and invertibility the initial condition should not matter asymptotically. One can re-write the model as $(1 - (\pi + \beta)L) Y_t = (1 - \pi L) e_t$. It is easy to see that if $\beta = 0$, then the parameter π is not identified. Assume that the model is point identified, so $\beta \neq 0$, but that identification is weak. This can be modeled as $\beta = \frac{C}{\sqrt{T}}$. If $K_T = diag(1/\sqrt{T}, 1)$, then:

$$K_T J_T(\theta_0) K_T \to^p \Sigma \quad \text{and} \quad K_T I_T(\theta_0) K_T \Rightarrow \Sigma + \begin{pmatrix} 0 & \xi \\ \xi & C\eta \end{pmatrix},$$

where Σ is a positive definite matrix while ξ and η are two Gaussian random variables (the derivation of this expression can be found in the Supplementary Appendix). That is, the difference between the two information matrices is asymptotically non-negligible compared with the information measure $J_T(\theta_0)$. □

Example 3 Another example is a weakly identified binary model. Assume that we observe an i.i.d. sample from the joint distribution of (Y_t, X_t), where

$$Y_t = \mathbb{I}\{Y_t^* > 0\}; \quad Y_t^* = \beta h(X_t, \pi) - U_t; \quad U_t | X_t \sim i.i.d. f(u).$$

Assume that the model is point-identified, and that a standard list of smoothness and moment existence conditions holds (see the Supplementary Appendix for details).

It is easy to see that if $\beta = 0$ then parameter π is unidentified. The weak identification embedding considered in Andrews and Cheng (2009) takes $\beta_0 = C/\sqrt{T}$. Again, for

$\theta = (\beta, \pi')'$ and $K_T = diag(1/\sqrt{T}, 1, \ldots, 1)$ we have that $K_T J_T(\theta_0) K_T$ converges in probability to a non-degenerate matrix, while $K_T(J_T(\theta_0) - I_T(\theta_0))K_T$ weakly converges to a non-zero random matrix with Gaussian entries:

$$K_T(J_T(\theta_0) - I_T(\theta_0))K_T \Rightarrow \begin{pmatrix} 0 & \xi' \\ \xi & C\eta \end{pmatrix},$$

where $(\xi', vec(\eta)')'$ is a Gaussian vector. □

The difference between the two measures of information can be used to construct a test to detect weak identification. A potential test should compare the two observed informations at the true parameter value. As argued in the introduction, however, tests of identification are less useful than weak identification-robust procedures so we do not pursue such tests here.

White (1982) shows in the context of quasi-MLE that the two measures of information may be asymptotically different if the likelihood is misspecified. As we point out above, even if the model is correctly specified the two informations may differ if identification is weak. While we are aware of one strand of the classical statistical literature which explores the difference between these different information measures, the literature on so-called non-ergodic models, these models are usually part of the LAMN (locally asymptotically mixed-normal) class, whereas the types of models which we consider in this paper are not in general LAMN.

3 Test for Full Parameter Vector

In this section, we suggest a test for a simple hypothesis on the full parameter vector, $H_0 : \theta = \theta_0$, which is robust to weak identification. To allow for the possibility of an embedding such as weak IV, we consider a so-called scheme of series. In a scheme of series we assume that we have a series of experiments indexed by the sample size: the data X_T of sample size T is generated by distribution $f_T(X_T; \theta_0)$, which may change as T grows. We assume that in the definition of all quantities in the previous section there is a silent index T. For example, the log-likelihood is $\ell_T(\theta) = \sum_{t=1}^{T} \log f_T(x_{T,t}|X_{T,t-1}; \theta)$, where the data is $X_T = (x_{T,1}, ..., x_{T,T})$ and $X_{T,t} = (x_{T,1}, ..., x_{T,t})$. All scores and information matrices also have this implied index T; for each fixed T the score $S_{T,t}$ is a process indexed by t, $S_{T,t}(\theta_0) = \frac{\partial}{\partial \theta'} \log f_T(X_{T,t}; \theta_0) = \sum_{j=1}^{t} s_{T,j}(\theta_0)$, and is a martingale with respect to

the sigma-field $\mathcal{F}_{T,t}$ generated by $X_{T,t}$. All other statistics are defined correspondingly. In this context, we introduce our first assumption:

Assumption 1 *Assume that there exists a sequence of constant matrices K_T such that:*

(a) for all $\delta > 0$, $\sum_{t=1}^{T} E\left(\|K_T s_{t,T}(\theta_0)\| \mathbb{I}\{\|K_T s_{t,T}(\theta_0)\| > \delta\} \middle| \mathcal{F}_{t-1}\right) \to 0$;

(b) $\sum_{t=1}^{T} K_T s_{t,T}(\theta_0) s_{t,T}(\theta_0)' K_T = K_T J_T(\theta_0) K_T \to^p \Sigma$, where Σ is constant positive-definite matrix.

Discussion of Assumption 1

Assumption 1(a) is a classical infinitesimality (or limit negligibility) condition. We can, if we prefer, replace it with a version of Linderberg's condition:

$$\sum_{t=1}^{T} E\left(\|K_T s_{t,T}\|^2 \mathbb{I}\{\|K_T s_{t,T}(\theta_0)\| > \delta\} \middle| \mathcal{F}_{t-1}\right) \to 0,$$

although this condition is stronger than 1(a). Assumption 1(b) imposes the ergodicity of the quadratic variation $J_T(\theta_0)$ of martingale $S_T(\theta_0)$, which rules out some potentially interesting models including persistent (unit root) processes and non-ergodic models.

Examples 1, 2 and 3 (cont.) Assumption 1 is trivially satisfied for the weak IV model, the ARMA (1,1) model with nearly canceling roots, and the weakly identified binary choice model (see the Supplementary Appendix for details).

Example 4 Assumption 1 can also be checked for an exponential family with weak identification. In particular, consider an exponential family with joint density of the form

$$f_T\left(X_t|\theta\right) = h(X_T) \exp\left\{\eta_T(\theta)' \sum_{t=1}^{T} H(x_t) - T A_T(\eta_T(\theta))\right\}. \tag{3}$$

Here, η is a $p-$dimensional reduced form parameter, while $\sum_{t=1}^{T} H(x_t)$ is a $p-$dimensional sufficient statistic. Model (3) covers VAR models with η being a set of reduced form VAR coefficients and $x_t = (Y_t', ..., Y_{t-p}')'$, where Y_t is a vector of data observed at time t, and the sufficient statistics are the sample autocovariances of the Y_t. Fernández-Villaverde et al. (2007) discuss the relationship between linearized DGSE models and VARs.

Suppose that we can partition the structural coefficient θ into sub-vectors α and β, $\theta = (\alpha', \beta')'$. We consider an embedding similar to that of Stock and Wright (2000) for

weak GMM, which we use to model β as weakly identified. In particular, we assume that

$$\eta_T(\theta) = m(\alpha) + \frac{1}{\sqrt{T}}\tilde{m}(\alpha, \beta),$$

where $\frac{\partial}{\partial \beta'} m(\alpha_0)$ and $\frac{\partial}{\partial \theta'}\tilde{m}(\alpha_0, \beta_0)$ are matrices of full rank ($\dim(\theta) = k = k_\alpha + k_\beta \leq p$). This means that while θ is identified for any fixed T, the likelihood is close to flat in directions corresponding to β. Assumption 1 is trivially satisfied for $K_T = \begin{pmatrix} \frac{1}{\sqrt{T}} Id_{k_\alpha} & 0 \\ 0 & Id_{k_\beta} \end{pmatrix}$ so long as the infinitesimality condition holds for the sequence $\left\{ \frac{1}{\sqrt{T}} H(x_t) \right\}_{t=1}^{T}$ and a law of large numbers holds for $H(x_t)H(x_t)'$ (i.e. $\frac{1}{T}\sum_{t=1}^{T} H(x_t)H(x_t)' \to^p E\left[H(x_t)H(x_t)'\right]$). □

The following theorem is a direct corollary of the multivariate martingale Central Limit Theorem (see Theorem 8, ch. 5 in Liptser and Shiryayev (1989)).

Theorem 1 *If Assumption 1 holds, then* $K_T S_T(\theta_0) \Rightarrow N(0, \Sigma)$, *and*

$$LM(\theta_0) = S_T(\theta_0) J_T(\theta_0)^{-1} S_T(\theta_0) \Rightarrow \chi^2_k, \tag{4}$$

where $k = dim(\theta_0)$.

Remark. There are a number of other ways to approach the problem of testing the full parameter vector. Since we consider a fully parametric model, so long as one only wishes to test hypotheses on the whole parameter vector, one could in principal obtain an exact test by simulating any statistic under the null. Alternatively, one could replace $\frac{J_T(\theta_0)}{T}$ with $\mathcal{I}(\theta_0) = E[J_T(\theta_0)/T]$ (Fisher's information) in the expression for the LM statistic. This would again produce an asymptotically χ^2_k statistic, but we contend that our original formulation is superior in many cases, since calculating $J_T(\theta_0)$ is much more straightforward than calculating $\mathcal{I}(\theta_0)$ when we do not have an analytic expression for the likelihood. In addition, if we weaken Assumption 1(b) to require only that Σ be an almost surely positive definite random matrix, then statement (4) still holds. In this sense, our formulation has the additional advantage of being robust to non-ergodicity, a characteristic not shared by the formulation using $\mathcal{I}(\theta_0)$. Statistical examples of non-ergodic models can be found in Basawa and Koul (1979).

Statement (4) of Theorem 1 suggests a test for simple hypotheses about the whole parameter vector θ. Unlike the classical ML Wald and LR tests, the derivation of the

asymptotic distribution of this statistic uses no assumptions about the strength of identification. The statistic is a special form of the classical LM (score) test, which is formulated as $LM = \frac{1}{T} S_T(\theta_0)' \hat{\mathcal{I}}^{-1} S_T(\theta_0)$, where $\hat{\mathcal{I}}$ is any consistent estimator of Fisher's information. Our suggested statistic plugs in $\frac{1}{T} J_T(\theta_0) = \frac{1}{T} [S(\theta_0)]_T$ for this estimator. It is important to note that while the true Fisher information is asymptotically degenerate under weak identification, the appropriately defined LM statistic (as in (4)) nevertheless achieves a χ^2 distribution asymptotically.

As already discussed, this test for the full parameter vector allows us to directly test the structural parameters in weakly identified exponential family models, including DSGE models which can be represented as VARs. In such models, our proposed test for the full parameter vector offers a number of advantages relative to other procedures in the literature. In particular, unlike the approach proposed by Guerron-Quintana, Inoue and Killian (2009), we require no assumptions on the strength of identification of the reduced form parameters. The test statistic is quite straightforward to compute, and maintains size well in simulation (see Section 7). Under strong identification, this test is asymptotically equivalent to the usual LM test, and thus inherits all of its properties. It is important to note, however, that the LM statistic calculated with other estimators of Fisher's information (for example $\frac{1}{T} I_T(\theta_0)$) is not necessarily robust to weak identification, as can be seen in the example of weak IV. It is also a bad idea to estimate the information matrix using an estimator of θ, i.e. to use $\frac{1}{T} J_T(\hat{\theta})$. All of these alternative formulations deliver asymptotically equivalent tests in strongly identified models, but this equivalence fails under weak identification.

4 Test for a Subset of Parameters

4.1 The Problem

In applied economics, it is very common to report separate confidence intervals for each one-dimensional sub-parameter in the (often quite multidimensional) parameter vector θ. Current standards require that each such confidence interval be valid, that is, it should have at least 95% coverage asymptotically (assuming the typical 95% confidence level). These one-dimensional confidence sets need not be valid jointly: if $\dim(\theta) = k$, the k-dimensional rectangle formed by the Cartesian product of the 1-dimensional confidence

intervals need not have 95% asymptotic coverage. Going the other direction, if one has a 95% confidence set for θ and projects it on the one-dimensional subspaces corresponding to the individual sub-parameters, the resulting confidence sets for the one-dimensional parameters will of course be valid. However, confidence sets obtained in such a manner (usually called the projection method) tend to be conservative.

Using our proposed test of the full parameter vector, which is robust to weak identification, we have the option to produce robust confidence sets for sub-parameters via the projection method. This approach has been used many times in the literature, for example by Dufour and Taamouti (2005) for weak IV and Guerron-Quintana, Inoue, and Killian (2009) for DSGE. The typical DSGE model has a large number of parameters to estimate (often between 20 and 60), which makes the projection method less attractive as the degree of conservativeness may be very high, which in turn makes the resulting confidence sets less informative.

For some intuition on the source of this conservativeness, imagine for a moment that we are concerned with a two-dimensional parameter $\theta = (\theta_1, \theta_2)'$, and have a t-statistic for each θ_i. Suppose, moreover, that these two statistics are asymptotically normal and asymptotically independent of each other. We can construct a confidence set for each parameter in two ways: the first and most commonly used is to invert the t-test for the corresponding sub-parameter, which is equivalent to using the the squared t-statistic and χ^2_1 critical values and yields $C_{1,\theta_i} = \left\{\theta_i : \frac{(\hat{\theta}_i - \theta_i)^2}{\sigma_i^2} \leq \chi^2_{1,.95}\right\}$. As an alternative, one may construct a joint confidence set for θ, which in this case will be an ellipse $C_{2,\theta} = \left\{\theta : \frac{(\hat{\theta}_1 - \theta_1)^2}{\sigma_1^2} + \frac{(\hat{\theta}_2 - \theta_2)^2}{\sigma_2^2} \leq \chi^2_{2,.95}\right\}$, and then use the projection method to obtain $C_{2,\theta_1} = \{\theta_1 : \exists \theta_2 \; s.t. \; (\theta_1, \theta_2)' \in C_{2,\theta}\}$ (and likewise for θ_2). One can notice that C_{2,θ_i} ultimately uses the same t-statistic as C_{1,θ_i}, but compares this statistic to the critical value of a χ^2_2 rather than a χ^2_1. As a result, in this example the projection method produces unnecessarily wide (and conservative) confidence sets for each sub-parameter.

The projection method, when applied to strongly identified models, produces a less powerful test than classical MLE. Thus, when using the projection method it is natural to combine it with a pre-test procedure which first discriminates between weakly and strongly identified models and then, based on the results of the test, uses either classical MLE or the projection method. There are two obstacles to such an approach: first, we are unaware of procedures for effectively discriminating between weak and strong identification in maximum likelihood. Second, the size properties of two-step testing

procedures are notoriously difficult to asses. Our approach is different, and instead constructs a test which maintains correct asymptotic size under weak identification, but which is equivalent to the classical MLE tests under strong identification.

We are aware of a number of papers dealing with this issue in the context of weak identification. In particular, Stock and Wright (2000) prove that for GMM, under some assumptions, if $\theta = (\alpha', \beta')'$ and α is well identified then it is possible to test the hypothesis $H_0 : \beta = \beta_0$ by comparing the GMM objective function, minimized with respect to α, to the critical values of a $\chi^2_{p-k_\alpha}$ distribution, where p is the number of moment conditions used and $k_\alpha = \dim(\alpha)$. Their result shows that it is possible to reduce the degrees of freedom for projection-based confidence sets in weak GMM provided the nuisance parameter is strongly identified.

Kleibergen and Mavroeidis (2009) prove that it is possible to extend this result to some models where the nuisance parameter may be weakly identified. They consider a test statistic, called $H(\theta_0)$ here, for testing the simple hypothesis $H_0 : \theta = \theta_0$ (they use the Anderson-Rubin and IV-LM tests). Assume again that $\theta = (\alpha', \beta')'$, and that the hypothesis of interest is $H_0 : \beta = \beta_0$. Kleibergen and Mavroeidis (2009) demonstrate that one can again use the quantiles of a $\chi^2_{p-k_\alpha}$ as critical values. This test is asymptotically similar if identification of the nuisance parameter is strong, and somewhat asymptotically conservative if identification of the nuisance parameter is weak. In this paper we consider a class of models which, as discussed above, differs from those in the weak GMM literature in that the variance of the moment conditions may be degenerate asymptotically, necessitating an alternative approach to eliminating nuisance parameters.

4.2 Classical LM Tests for Composite Hypotheses

We assume that $\theta = (\alpha', \beta')'$. We are interested in testing the composite hypothesis $H_0 : \beta = \beta_0$, treating α as a nuisance parameter. The classical theory for maximum likelihood considers two LM tests for such a setting: Rao's score test and Neyman's $C(\alpha)$-test.

Let $S_T(\theta) = (S_\alpha(\theta)', S_\beta(\theta)')'$, $\mathcal{I}(\theta) = \begin{pmatrix} \mathcal{I}_{\alpha\alpha} & \mathcal{I}_{\alpha\beta} \\ \mathcal{I}'_{\alpha\beta} & \mathcal{I}_{\beta\beta} \end{pmatrix}$ be Fisher's information, and $\hat{\theta}_0$ be the restricted ML estimator of θ, under the restriction $\hat{\beta} = \beta_0$. Assume, in addition, that all martingales introduced in Section 2 are divided into sub-matrices corresponding

to α and β. Rao's score test is based on the statistic $Rao = \frac{1}{T} S_T(\hat{\theta}_0)' \mathcal{I}(\hat{\theta}_0)^{-1} S_T(\hat{\theta}_0)$.

Neyman's $C(\alpha)$ test was developed as a locally asymptotically most powerful (LAMP) test for composite hypotheses in the classical ML model (see Akritas (1988)). The statistic is defined as

$$C(\alpha) = \frac{1}{T} \left(S_\beta - \mathcal{I}'_{\alpha\beta} \mathcal{I}^{-1}_{\alpha\alpha} S_\alpha \right)' \mathcal{I}^{-1}_{\beta\beta,\alpha} \left(S_\beta - \mathcal{I}'_{\alpha\beta} \mathcal{I}^{-1}_{\alpha\alpha} S_\alpha \right) \Big|_{\theta=(\hat{\alpha},\beta_0)},$$

where $\hat{\alpha}$ is any $\sqrt{T}$ consistent estimator of α, and $\mathcal{I}_{\beta\beta,\alpha} = \mathcal{I}_{\beta\beta} - \mathcal{I}_{\beta\alpha} \mathcal{I}^{-1}_{\alpha\alpha} \mathcal{I}_{\alpha\beta}$.

Kocherlakota and Kocherlakota (1991) show that the two statistics are the same asymptotically if one takes $\hat{\alpha}$ in Neyman's $C(\alpha)$ test to be the restricted MLE. If the classical ML assumptions are satisfied then both statistics are distributed $\chi^2_{k_\beta}$ asymptotically. In this paper, we suggest a statistic which is asymptotically equivalent to both Rao's score and Neyman's $C(\alpha)$ if the classical ML assumptions are satisfied. In particular, we consider the same LM statistic defined in (4) but evaluated at $\theta = (\hat{\alpha}, \beta_0)$, where $\hat{\alpha}$ is the restricted MLE, that is, the solution to equation $S_\alpha(\hat{\alpha}, \beta_0) = 0$. One can easily see that

$$\widetilde{LM}(\beta_0) = LM(\hat{\alpha}, \beta_0) = S'_\beta \left(J_{\beta\beta} - J_{\beta\alpha} J^{-1}_{\alpha\alpha} J'_{\beta\alpha} \right)^{-1} S_\beta \Big|_{\theta=(\hat{\alpha},\beta_0)}. \tag{5}$$

5 Test for a Subset of Parameters- Strong Identification

In this section, we establish that if the nuisance parameter α satisfies conditions implying the asymptotic normality of its restricted MLE then the statistic defined in (5) has a $\chi^2_{k_\beta}$ distribution asymptotically regardless of the strength of identification of β. One implication of these results is that when α is strongly identified, our proposed subset test has a $\chi^2_{k_\beta}$ distribution asymptotically.

5.1 Asymptotic Normality of $\hat{\alpha}$

When we test $H_0 : \beta = \beta_0$, under the null α is the only unknown parameter. Below, we provide conditions which guarantee that the restricted maximum likelihood estimator of α will be asymptotically normal. We adapt Baht's (1974) result on the consistency

and asymptotic normality of the MLE for time series. We call α strongly identified if, in addition to the conditions below, information about α goes to infinity as the sample size grows. Let $A_{\alpha\alpha,T} = J_{\alpha\alpha,T} - I_{\alpha\alpha,T}$, where the last two quantities are the sub-matrices of $J_T(\theta_0)$ and $I_T(\theta_0)$ corresponding to α.

Assumption 2 *Assume that matrix K_T from Assumption 1 is diagonal and $K_{\alpha,T}$ and $K_{\beta,T}$ are the sub-matrices of K_T corresponding to α and β, respectively.*

(a) $K_{\alpha,T}A_{\alpha\alpha,T}K_{\alpha,T} \to^p 0$;

(b) for any $\delta > 0$ we have

$$\sup_{\|K_{\alpha,T}^{-1}(\alpha_1-\alpha_0)\|<\delta} \|K_{\alpha,T}(I_{\alpha\alpha}(\alpha_1,\beta_0) - I_{\alpha\alpha}(\alpha_0,\beta_0))K_{\alpha,T}\| \to^p 0.$$

Lemma 1 *If Assumptions 1 and 2 are satisfied, then*

$$K_{\alpha,T}^{-1}(\hat{\alpha} - \alpha_0) = K_{\alpha,T}^{-1}J_{\alpha\alpha,T}^{-1}S_{\alpha,T} + o_p(1) \Rightarrow N(0, \Sigma_{\alpha\alpha}^{-1}). \tag{6}$$

Discussion of Assumption 2. Assumption 2(a) may be formulated as $J_{\alpha\alpha,T}^{-1}I_{\alpha\alpha,T} \to^p Id_{k_\alpha}$, which requires that the two information matrices be the same asymptotically. We mentioned a condition of this nature in our discussion of weak identification in Section 2. One approach to checking 2(a) in many contexts is to establish a Law of Large Numbers for $A_{\alpha\alpha,T}$. Indeed, $A_{\alpha\alpha,T}$ is a martingale of the form

$$A_{\alpha\alpha,T} = \sum_{t=1}^{T} \frac{1}{f(x_t|X_{t-1},\theta_0)} \frac{\partial^2}{\partial\alpha\partial\alpha'} f(x_t|X_{t-1},\theta_0).$$

If the terms $\frac{1}{f(x_t|X_{t-1},\theta_0)} \frac{\partial^2}{\partial\alpha\partial\alpha'} f(x_t|X_{t-1},\theta_0)$ are uniformly integrable and $K_{\alpha,T}$ converges to zero no slower than $\frac{1}{\sqrt{T}}$, then the martingale Law of Large Numbers gives us Assumption 2(a).

Assumption 2(b) is an assumption on the smoothness of the log-likelihood. We can reformulate it using the third derivatives:

$$\Lambda_{\alpha\alpha\alpha_i,T}(\theta) = \sum_{t=1}^{T} \frac{1}{f(x_t|X_{t-1},\theta)} \frac{\partial^3}{\partial\alpha_i\partial\alpha\partial\alpha'} f(x_t|X_{t-1},\theta). \tag{7}$$

An alternative to Assumption 2(b) is:

Assumption 2 *(b')* for any i: $K_{\alpha_i,T} \sup_{\|K_{\alpha,T}^{-1}(\alpha-\alpha_0)\|<\delta} \|K_{\alpha,T}\Lambda_{\alpha_i\alpha\alpha,T}K_{\alpha,T}\| \to^p 0$.

Lemma 2 *Assumptions 1, 2(a) and 2(b') imply assumption 2(b).*

If we have $K_{\alpha,T} = 1/\sqrt{T}Id_{k_\alpha}$, as is often the case for strongly identified α, then Assumption 2(b') usually holds due to the Law of Large Numbers since the normalization is excessive.

Strong Identification If in addition to Assumption 2 we assume that $K_{\alpha,T} \to 0$, it is clear that $\hat{\alpha}$ will be consistent for α_0. In such instances we say that α is strongly identified, as information about α goes to infinity as the sample grows and the MLE for α is consistent and asymptotically normal (under the null hypothesis $\beta = \beta_0$).

5.2 Result

As we show in Section 3, to test a simple hypothesis about the whole parameter vector it is enough to have a CLT for the score function. Kleibergen and Mavroeidis (2009) impose a stronger assumption for the their test of a subset of parameters, namely that the CLT also hold for the derivative of the moment condition (in fact, they impose a functional CLT). For our test of a subset of parameters, we likewise need an additional assumption, specifically a CLT on the derivative of the score, which is directly related to the martingale A_T (the difference of the two information matrices).

Assumption 3 *Consider the sequence of martingales* $M_T = (S_T(\theta_0)', vec(A_{\alpha,\beta,T}(\theta_0))')' = \sum_{t=1}^T m_{t,T}$. *Assume that there exists a sequence of non-stochastic diagonal matrices* $K_{M,T}$ *such that:*

(a) for all $\delta > 0$, $\sum_{t=1}^T E\left(\|K_{M,T}m_{t,T}\|\mathbb{I}\{\|K_{M,T}m_{t,T}\| > \delta\}|\,\mathcal{F}_{t-1}\right) \to 0$;

(b) $\sum_{t=1}^T K_{M,T}m_{t,T}m_{t,T}'K_{M,T} \to^p \Sigma_M$, *where* Σ_M *is a constant matrix whose sub-matrix* Σ *corresponding to the martingale* S_T *is positive definite.*

Let us define the martingales associated with the third derivative of the likelihood function:

$$\Lambda_{\alpha_i\alpha_j\beta_n} = \sum_{t=1}^T \frac{1}{f(x_t|X_{t-1},\theta_0)} \cdot \frac{\partial^3 f(x_t|X_{t-1},\theta_0)}{\partial\alpha_i\partial\alpha_j\partial\beta_n}.$$

If we can interchange integration and differentiation three times then each entry of $\Lambda_{\alpha\alpha\beta,T}$ is a martingale. For the proof of the theorem below we will also need the following assumptions:

Assumption 4 *(a)* $\lim_{T\to\infty} K_{\alpha_i,T}K^{-1}_{\alpha_i\beta_j,T}K_{\beta_j,T} = C_{ij}$, *where* C *is some finite matrix (which may be zero).*

(b) $K_{\alpha_i,T}K_{\alpha_j,T}K_{\beta_n,T}\sqrt{[\Lambda_{\alpha_i\alpha_j\beta_n}]} \to^p 0$ *for any* i,j,n.

(c) $\sup_{\|K^{-1}_{\alpha,T}(\alpha-\alpha_0)\|<\delta} \left\| K_{\beta_j,T}K_{\alpha,T}(\frac{\partial}{\partial\beta_j}I_{\alpha\alpha}(\alpha,\beta_0) - \frac{\partial}{\partial\beta_j}I_{\alpha\alpha}(\alpha_0,\beta_0))K_{\alpha,T} \right\| \to^p 0.$

Discussion of Assumption 4

Assumptions 4(b) and (c) state that the higher order derivatives with respect to α are not important for the analysis. If α is strongly identified, then Assumptions 4(b) and (c) generally hold, and can be checked using some Law of Large Numbers, since the normalization $K^2_{\alpha,T}$ or $K^3_{\alpha,T}$ converges to zero very quickly. Finally, Assumption 4 holds trivially for weak IV, as well as for the exponential family case discussed in section 3.

Theorem 2 *If Assumptions 2, 3 and 4 are satisfied then under the null* $H_0 : \beta = \beta_0$ *we have* $\widetilde{LM}(\beta_0) \Rightarrow \chi^2_{k_\beta}$.

Examples 1, 2 and 3 (cont.) Assumptions 2, 3 and 4 trivially hold for the weak IV model when we test the composite hypothesis $H_0 : \beta = \beta_0$. The resulting test is the K-test introduced in Kleibergen (2002) and Moreira (2001). In the Supplementary Appendix, we show that Assumptions 2, 3 and 4 hold in the ARMA(1,1) model with nearly canceling roots and the weakly identified binary choice model for testing a hypothesis $H_0 : \pi = \pi_0$ about the weakly-identified parameter π. Thus, our subset test for this parameter is robust to weak identification. In Section 7 we show that the finite sample properties of the test are remarkably good in the ARMA(1,1) model.

5.3 How Our Result Differs from the Previous Literature

As discussed above, Stock and Wright (2000) develop a framework for weakly identified GMM and construct a test for the hypothesis $H_0 : \beta = \beta_0$ when the nuisance parameter α is strongly identified (Theorem 3 in Stock and Wright (2000)). They consider GMM

with moment condition $Em(x_t, \alpha, \beta) = 0$ and construct a statistic based on

$$S(\theta) = (\frac{1}{\sqrt{T}} \sum_{t=1}^{T} m(x_t; \theta))' W_T^{-1}(\theta) (\frac{1}{\sqrt{T}} \sum_{t=1}^{T} m(x_t; \theta)),$$

where $W_T(\theta)$ is a consistent estimator of the variance of the moment condition. They show that, for $\hat{\alpha} = \arg\min_\alpha S(\alpha, \beta_0)$, their statistic $S(\hat{\alpha}, \beta_0)$ has an asymptotic χ^2 distribution with degrees of freedom equal to $p - k_\alpha$, where $p = \dim(m(x_t, \theta))$ and $k_\alpha = \dim(\alpha)$.

Kleibergen (2005) considers an alternative statistic based on the LM test for GMM and proves that this statistic, minimized over α, is also the basis of a valid test of $H_0 : \beta = \beta_0$ when α is strongly identified. In our context, however, if we use the score of the log-likelihood as the GMM moment condition the system is just-identified and Kleibergen's KLM statistic is equal to Stock and Wright's S statistic.

Our result, though of a similar flavor, is quite different and is not covered by these previous results. First, the weak ML model does not satisfy the assumptions in the above mentioned papers. Specifically, if we consider ML estimation as GMM using the moment condition $ES_T(\theta_0) = 0$, the variance matrix of our moment condition (information matrix) is directly linked to identification. In particular, the matrix $W_T(\theta)$ (to use Stock and Wright's notation) becomes degenerate asymptotically, which is ruled out by the assumptions of Stock and Wright (2000), Kleibergen (2005), and Kleibergen and Mavroeidis (2009). Second, we apply a different principle to go from a test of the full parameter vector to a test for a subset of parameters. In the above mentioned papers the authors minimize the statistic over the nuisance parameter, while we plug in the restricted MLE. In fact, in our context minimizing the statistic over the nuisance parameter does not necessarily lead to a χ^2 distribution, as illustrated in the following weak IV example.

Example 1 (cont.) Let us return to the weak IV model and consider the LM statistic for $LM(\beta, \pi)$ for testing the whole parameter vector $\theta = (\beta, \pi')'$, defined as in equation (4). Suppose we wish to test the composite hypothesis $H_0 : \beta = \beta_0$ by considering the concentrated statistic:

$$LM^c(\beta_0) = \min_\pi LM(\beta_0, \pi) = LM(\beta_0, \tilde{\pi}).$$

We can show (see proof in the Supplementary Appendix) that

$$LM^c(\beta_0) = \frac{(Q_S + Q_T) - \sqrt{(Q_S + Q_T)^2 - 4Q_{ST}^2}}{2},$$

where Q_S, Q_T, and Q_{ST} are defined as in Andrews, Moreira, and Stock (2006). If the instruments are weak, that is if $\pi = C/\sqrt{T}$, then the asymptotic distribution of $LM^c(\beta_0)$ is stochastically dominated by a χ_1^2, and the resulting test is conservative.

6 Test for a Subset of Parameters- Weak Identification

In the previous section we show that our subset-test statistic $\widetilde{LM}(\beta_0)$ for the composite hypothesis $H_0 : \beta = \beta_0$ is asymptotically $\chi^2_{k_\beta}$ when the nuisance parameter α is strongly identified, without any assumptions about the identification of β. Strong identification, however, is not necessary for the validity of our proposed test statistic. Below, we present two examples in which the nuisance parameter is weakly identified but $\widetilde{LM}(\beta_0)$ nonetheless has a $\chi^2_{k_\beta}$ distribution asymptotically.

6.1 Weak IV Case

Example 1(cont.) Here we consider a weak IV model with one endogenous variable, when the hypothesis tested is one about π, that is, $H_0 : \pi = \pi_0$, while the weakly identified parameter β is treated as a nuisance parameter. For simplicity we consider a slightly different version of the quadratic variation of S, namely the *expected quadratic variation*.

$$\widetilde{J} = \langle S \rangle = \sum_{t=1}^{T} E\left(s_t s_t' | \mathcal{F}_{t-1}\right) = \begin{pmatrix} \pi' Z' Z \pi & \beta \pi' Z' Z \\ \beta Z' Z \pi & (1+\beta^2) Z' Z \end{pmatrix}.$$

The difference between J_T and $\widetilde{J}$ doesn't matter asymptotically as $J_T^{-1}\widetilde{J} \to^p Id_{k+1}$ uniformly over the strength of instruments.

According equation (5) our statistic of interest is $\widetilde{LM}(\pi_0) = LM(\hat{\beta}, \pi_0)$, where $\hat{\beta}$ is the restricted ML estimator of β, and $LM(\beta, \pi_0)$ is defined as in (4) with the slight modification that $\widetilde{J}$ is used in place of J. Note that $S_\beta(\hat{\beta}, \pi_0) = 0$, so we can explicitly solve for $\hat{\beta}$ as $\hat{\beta} = \frac{\pi_0' Z' Y}{\pi_0' Z' Z \pi_0}$. Simple calculations show that

$$LM\left(\hat{\beta}, \pi_0\right) = \left(\hat{\beta}\hat{U} + V_0\right)' Z \left((1+\hat{\beta}^2) Z'Z - \frac{\hat{\beta}^2 Z'Z\pi_0\pi_0' Z'Z}{\pi_0' Z'Z\pi_0}\right)^{-1} Z' \left(\hat{\beta}\hat{U} + V_0\right),$$

where $\hat{U} = Y - \hat{\beta}Z\pi_0$.

Lemma 3 *If $\pi_0 = C/\sqrt{T}$, we have $LM(\hat{\beta}, \pi_0) \Rightarrow \chi^2_k$.*

The idea of the proof is the following. Under the weak instruments embedding, $\hat{\beta}$ is not consistent but is asymptotically normal. We can show that $(Z'Z)^{-1/2}Z'\hat{U}, \hat{\beta}$ and $(Z'Z)^{-1/2}Z'V_0$ are asymptotically normal and asymptotically uncorrelated with each other. If we consider statistic $LM(\hat{\beta}, \pi_0)$, conditional on $\hat{\beta}$ it becomes a correctly normalized quadratic form of an asymptotically normal $k-$dimensional random variable and thus conditionally asymptotically χ^2_k. As a result, unconditional convergence holds as well.

6.2 Case Where Score is Linear in α

The case considered in the previous subsection is interesting in that the nuisance parameter is weakly identified, but is somewhat trivial since the parameter tested is strongly identified. We can to a limited extent generalize this result to more interesting contexts. Below, we consider the problem of testing a hypothesis about a weakly identified parameter in an exponential family model. The nuisance parameter will be divided into two subsets, one of which is strongly identified while the other is weakly identified. We will make the very restrictive assumption that the weakly identified nuisance parameters enter linearly.

Example 4 (cont.) Assume that the experiment at time T is generated by the exponential family (3). As already discussed, model (3) covers VAR models, and many linearized DGSE models can be represented as VARs (see Fernández-Villaverde et al. (2007)). Assume that we are interested in structural parameters $\theta = (\alpha_1', \alpha_2', \beta')'$, where the relation between the structural and reduced form parameters is given by

$$\eta_T(\theta) = m(\alpha_1) + \frac{1}{\sqrt{T}} n(\alpha_1, \beta)\alpha_2 + \frac{1}{\sqrt{T}} r(\alpha_1, \beta). \tag{8}$$

We assume that the matrix $\left(\frac{\partial}{\partial \alpha_1'} m(\alpha_1), n(\alpha_1, \beta), \frac{\partial}{\partial \beta'} n(\alpha_1, \beta)\alpha_2 + \frac{\partial}{\partial \beta'} r(\alpha_1, \beta)\right)$ has full rank $k = dim(\theta) \leq p$ and call this the *rank assumption*. That is, we assume that the structural parameters are identified, though only α_1 is strongly identified (parameters α_2 and β are weakly identified). We are interested in testing a composite hypothesis

$H_0 : \beta = \beta_0$, treating $\alpha = (\alpha_1', \alpha_2')'$ as a nuisance parameter. We use the $\widetilde{LM}(\beta_0)$ statistic defined in (5).

Theorem 3 *Assume that in the model defined by equations (3) and (8) which satisfies the rank assumption the following convergence holds at the true value of* θ_0*:*

(a) $A_T(\eta) \to A(\eta)$*, as* $T \to \infty$ *in a neighborhood of* η_∞ *and the first four derivatives of* A_T *at* η_∞ *converge to those of* $A(\cdot)$*;*

(b) $\frac{1}{T}\sum_{t=1}^{T} H(x_t) \to^p \dot{A}$*;*

(c) $\frac{1}{T}\sum_{t=1}^{T}\left(H(x_t) - \dot{A}\right)\left(H(x_t) - \dot{A}\right)' \to^p -\frac{\partial^2}{\partial\eta\partial\eta'}A(\eta_\infty) = -\ddot{A}$*, where* $\ddot{A}$ *is a positive-definite matrix;*

(d) $\frac{1}{T}\sum_t H_i(x_t)H(x_t)H(x_t) = O_p(1)$ *for any* i*.*

Then under the null we have $\widetilde{LM}(\beta_0) \Rightarrow \chi^2_{k_\beta}$*.*

7 Simulation Results

We have a number of simulation results which both support our theoretical results and suggest directions for further research. We focus on simulation results from three models: a simple DSGE model based on Clarida, Gali, and Gertler (1999), a nonlinear extension of the standard weak IV model discussed earlier in this paper, and the ARMA(1,1) model with nearly canceling roots. In all cases, we simulate the behavior of our proposed statistics and compare the finite sample distributions of the statistics in question to their limiting distributions. In the DSGE example, we argue that estimation in the model behaves in a manner consistent with weak identification, and that our proposed statistics offer a substantial improvement over the usual Wald-based statistics for testing in this model. For the other two models, we use a standard specification for weak identification and show that our proposed tests have good properties in simulations.

7.1 DSGE Model

We consider a simple DSGE model based on Clarida, Gali and Gertler (1999). For this model, we first explore the properties of the ML estimator and the usual ML-based test statistics, then discuss the properties of the information matrix, and finally explore the

behavior of our proposed test statistics, both for the full parameter vector and for subsets of parameters.

The (log-linearized) equilibrium conditions for the model are

$$\begin{aligned}
\beta E_t \pi_{t+1} + \kappa x_t - \pi_t + \varepsilon_t &= 0, \\
-[r_t - E_t \pi_{t+1} - rr_t^*] + E_t x_{t+1} - x_t &= 0, \\
\alpha r_{t-1} + (1-\alpha)\phi_\pi \pi_t + (1-\alpha)\phi_x x_t + u_t &= r_t, \\
rr_t^* &= \rho \Delta a_t,
\end{aligned}$$

while the exogenous variables (Δa_t and u_t) evolve according to

$$\begin{aligned}
\Delta a_t = \rho \Delta a_{t-1} + \varepsilon_{a,t}; \quad u_t = \delta u_{t-1} + \varepsilon_{u,t}; \\
(\varepsilon_t, \varepsilon_{a,t}, \varepsilon_{u,t})' \sim iid\, N(0, \Sigma); \Sigma = diag(\sigma^2, \sigma_a^2, \sigma_u^2).
\end{aligned}$$

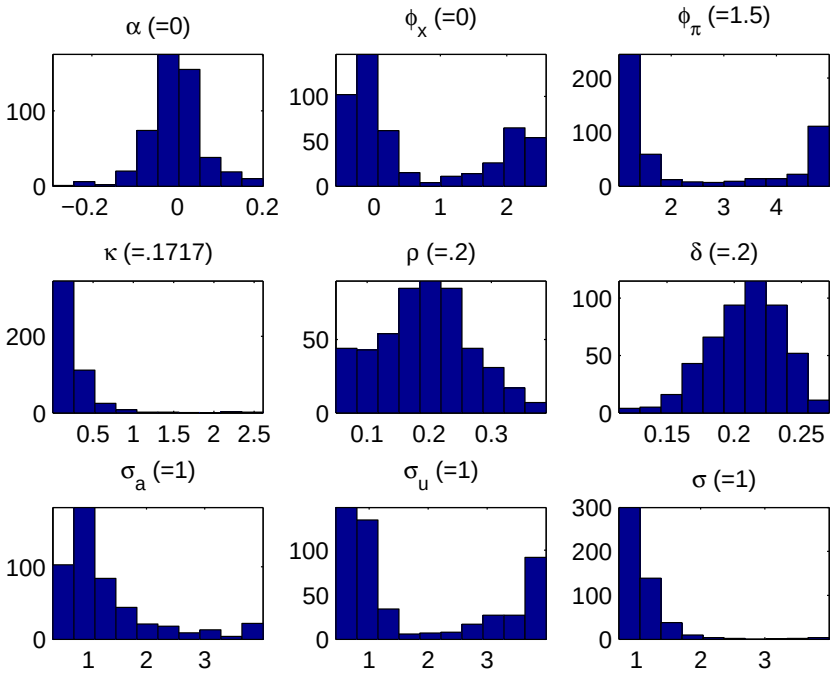
The model has ten parameters: the discount rate β, the structural parameters κ, ϕ_x, ϕ_π, and α, and the parameters describing the evolution of the exogenous variables. We calibrate the structural parameters at generally accepted values: $\beta = .99$, $\kappa = \frac{(1-\theta)(1+\phi)(1-\beta\theta)}{\theta} \approx .1717$, $\phi_x = 0$, $\phi_\pi = 1.5$ and $\alpha = 0$. For the parameters describing the exogenous variables, we choose $\rho = .2$ and $\delta = .2$, to introduce a degree of persistence while maintaining stationarity, and set $\sigma_a = 1$, $\sigma_u = 1$, and $\sigma = 1$. Using this model, we generate samples of size 300 and then discard the first 100 observations. We use only the last 200 observations from each simulation draw for the remainder of the analysis. Given well-documented problems with estimating β in many models, from this point forward we also calibrate this parameter at its true value, and conduct the analysis using the remaining 9 parameters.[5]

7.1.1 MLE Monte-Carlo Results

We begin by examining the behavior of the maximum likelihood estimator for the nine non-calibrated parameters in the model. We report histograms for the resulting estimates in Figure 1 (based on 500 Monte-Carlo draws). As can be seen from the figure, the distribution of many of the estimates is quite far from the normal limiting distribution of the maximum likelihood estimator under the usual assumptions. Moreover, it appears that this non-normality is not purely the result of bad behavior on the part of one parameter: after experimenting with calibrating (to their true values) a number of different

[5] We conducted extensive simulations, only some of which are presented here. Additional results are available from the authors by request.

Figure 1: Histogram of the unrestricted ML parameter estimates. The true value for each parameter is given in parenthesis at the top of its subplot.



parameters, it appears that we need to calibrate at least three parameters before the distributions of the remaining parameters begin to appear well-approximated by normal distributions.

Table 1: Size of Classical ML Tests for the 9-dimensional hypothesis $H_0 : \theta = \theta_0$.

	LR	Wald $(I(\theta_0))$	Wald $(I(\hat{\theta}))$	Wald $(J(\theta_0))$	Wald $(J(\hat{\theta}))$	$LM^*(\theta_0)$
Size of 5% Test	3.20%	65.45%	63.05%	68.05%	68.15%	6.55%
Size of 10% Test	7.05%	67.20%	64.30%	70.80%	71.00%	8.60%

While the results in Figure 1 show that the usual asymptotics for the ML estimator provide a poor approximation to its finite-sample distribution in this model, our theoretical results focus on questions of inference rather than estimation, so we also look at the behavior of the usual maximum likelihood tests for this model. We consider each of the trinity of classical tests (LR, Wald, and LM) in turn, focusing on tests of the full parameter vector. Specifically, we test the hypothesis $H_0 : \theta = \theta_0$, where θ is the vector consisting of all parameters other than β, and θ_0 is its true value. Under the usual assumptions for ML, all of these statistics should have a χ^2_9 distribution asymptotically. In simulations, however, the distribution of these statistics appears quite far from a χ^2_9.

To illustrate this fact, in Table 1 we list the size of a number of classical test statistics which, under classical assumptions, should have asymptotic size 5% or 10% (for the left and right columns, respectively, based on 2000 simulations). These sizes were generated by calculating the appropriate test statistic in simulation and comparing it to the 95th (or 90th) percentile of a χ^2_9 distribution. The LM statistic listed in Table 1 is calculated as $LM^*(\theta_0) = S(\theta_0)' I_T^{-1}(\theta_0) S(\theta_0)$ where $I(\theta_0) = -\ddot{\ell}(\theta_0)$ is the observed information (rather than with $J_T(\theta_0)$ as our LM statistic, $LM(\theta_0) = S(\theta_0)' J_T^{-1}(\theta_0) S(\theta_0)$). Table 1 also lists four variations on the Wald statistic, corresponding to different estimators of the asymptotic variance used in $(\hat{\theta} - \theta_0)\hat{V}^{-1}(\hat{\theta} - \theta_0)$. In particular, Wald $(I(\hat{\theta}))$ is the usual Wald statistic which uses the inverse of the observed information, evaluated at $\hat{\theta}$, to estimate the asymptotic variance. Wald $(I(\theta_0))$, on the other hand, evaluates the observed information at the true parameter value. Likewise, Wald $(J(\hat{\theta}))$ and Wald $(J(\theta_0))$ use J_T^{-1} as the estimator of the asymptotic variance, calculated at $\hat{\theta}$ and θ_0 respectively.

As can be seen in Table 1, the LR statistic is conservative. All versions of the Wald statistic which we consider severely overreject. Finally, the usual LM statistic (calculated using the negative hessian) somewhat overrejects at the 5% level and underrejects at the 10% level; however, additional simulation results show that the empirical distribution of this LM statistic is extremely poorly approximated by a χ^2_9, and that the seemingly small size distortions reported in Table 1 are entirely due to the the fact that the two cdfs cross near the 7% level. Taken together, these results strongly suggest that the usual approaches to ML estimation and inference are poorly behaved when applied to this model.

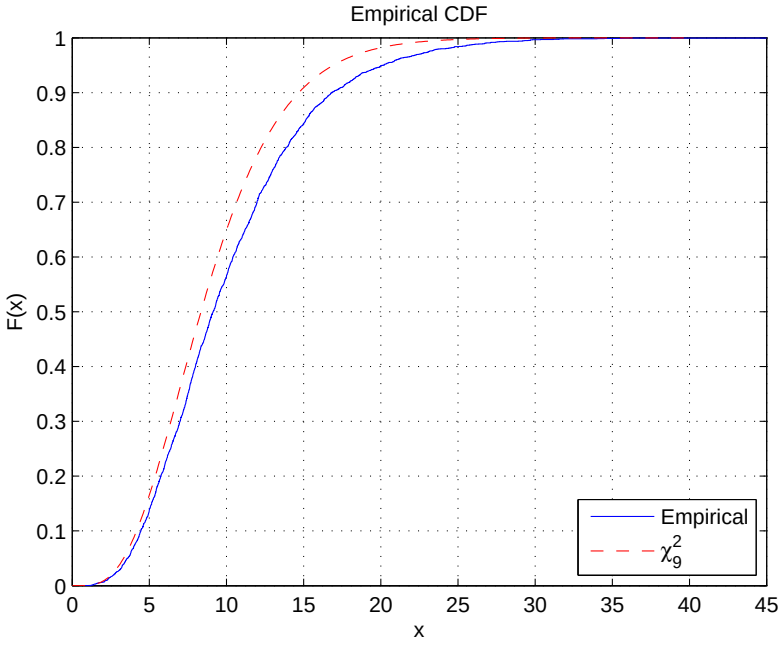
7.1.2 Behavior of the Information Matrix

Having examined the behavior of the usual ML estimator and tests in this model, we can also look directly at the properties of the information matrix. In Section 2 we associated weak identification with the difference between two information measures $A_T(\theta_0)$ being large compared to $J_T(\theta_0)$. We point out that observed incremental information $J_T(\theta_0)$ is an almost surely positive-definite matrix by construction, while $A_T(\theta_0)$ is a mean zero random matrix. If $A_T(\theta_0)$ is negligible compared to $J_T(\theta_0)$, then the observed information $I_T(\theta_0) = J_T(\theta_0) - A_T(\theta_0)$ is positive definite for the majority of realizations. We can check positive-definiteness of $I_T(\theta_0)$ directly in simulations. Considering the observed

information evaluated at the true value ($I_T(\theta_0) = -\frac{\partial^2}{\partial\theta\partial\theta'}\ell(\theta_0)$), we see that it has at least one negative eigenvalue in over 95% of simulation draws, and at least two negative eigenvalues in over 40% of simulation draws (based on 2000 simulations). While this falls far short of a formal test for weak identification, it is consistent with the idea that weak identification is the source of the bad behavior of ML estimation in this model.

7.1.3 LM Test for Full Parameter Vector

Figure 2: CDF of simulated LM statistic introduced in Theorem 1 compared to χ^2_9



We now turn to the weak identification-robust statistics discussed earlier in this paper. We begin by considering the behavior of the the test for the full parameter vector described in Section 3. As the reader will recall, under appropriate assumptions we have that $LM(\theta_0) \Rightarrow \chi^2_k$ under $H_0 : \theta = \theta_0$, where $LM(\theta)$ is defined in (4). In Figure 2, we plot the CDF of the simulated distribution of $LM(\theta_0)$, together with a χ^2_9. If we use χ^2_9 critical values to construct a test based on this statistic, a 5% test rejects 9.84% of the time, while a 10% test rejects 16.68% of the time: though this shows that the test based on $LM(\theta_0)$ and χ^2_9 critical values is not exact, the χ^2 approximation is far better for $LM(\theta_0)$ than for the usual Wald or LM statistics.

Table 2: Simulated size of a test $H_0 : b = b_0$ treating vector a as a nuisance parameter. In both cases the tested parameter is 6-dimensional. Statistic $\widetilde{LM}(b_0)$ is defined in equation (5) and statistic $LM(a, b)$ is defined in (4).

	$b = (\phi_x, \phi_\pi, \kappa, \sigma_a, \sigma_u, \sigma)$, $a = (\alpha, \rho, \delta)$		$b = (\alpha, \rho, \delta, \sigma_a, \sigma_u, \sigma)$ $a = (\phi_x, \phi_\pi, \kappa)$	
Test Statistic	5%	10%	5%	10%
$\widetilde{LM}(b_0)$	7.99%	15.28%	8.95%	15.40%
$\min_a LM(a, b_0)$	6.41%	12.99%	7.50%	13.30%

7.1.4 Subset Tests

Finally, we simulate tests for subsets of parameters. Specifically, as before we consider a partition of the parameter vector, $\theta = (a', b')'$, and consider the problem of testing $H_0 : b = b_0$ without any restrictions on a. In this context, we simulate two tests. One is based on the LM statistic evaluated at $(\hat{a}, b_0)$ for $\hat{a}$ the restricted ML estimator, which we have discussed extensively in this paper. The other is based on $\min_a LM(a, b_0)$, suggested Stock and Wright (2000) for GMM when a is strongly identified. If the results of Kleibergen and Mavroeidis (2009) can be extended to the current DSGE setting, when a is weakly identified the asymptotic distribution of this statistic will be dominated by that of a $\chi^2_{k_b}$. For both approaches, and for several subsets of parameters, we simulate the distribution of the statistic and then construct tests using quantiles from the $\chi^2_{k_b}$ distribution as critical values.

We first consider[6] testing the six parameters other than α, ρ, and δ, (so we have $a = (\alpha, \rho, \delta)$ and $b = (\phi_x, \phi_\pi, \kappa, \sigma_a, \sigma_u, \sigma)$). The size of 5% and 10% tests based on these statistics using asymptotic (χ^2_6) critical values is given in Table 2. As can be seen, while the χ^2_6 distribution does not provide a perfect approximation to the distribution of either statistic, it is fairly close. Both statistics tend to over-reject, so since the test based on $\min_a LM(a, b_0)$ is more conservative by construction it performs somewhat better.

We next consider testing the six parameters other than ϕ_x, ϕ_π and κ (so $a = (\phi_x, \phi_\pi, \kappa)$, while $b = (\alpha, \rho, \delta, \sigma_a, \sigma_u, \sigma)$). Again, the tests slightly over-reject compared to their asymptotic size.

Finally, we may be interested in testing only one parameter at a time (for example to generate confidence sets). Based on 1000 simulations, we report test sizes for each parameter separately in Table 3. The results for $LM(\hat{\alpha}, \beta_0)$ are similar to those in the other parameter subsets, although the degree of over-rejection is larger for most

[6] Additional simulation results are available upon request.

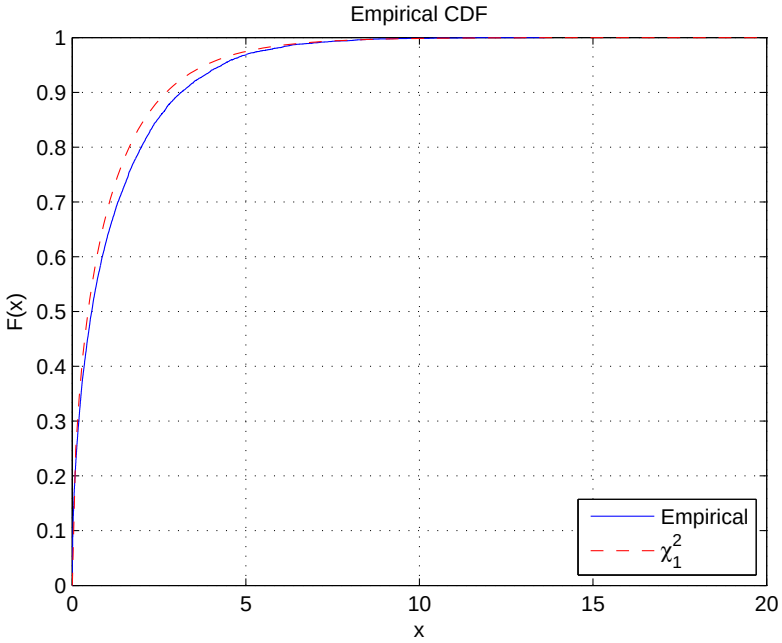
Table 3: Simulated Test Size for one dimensional hypotheses about each parameter separately

Parameter	$\widetilde{LM}(\beta_0)$ 5%	$\widetilde{LM}(\beta_0)$ 10%	$\min_\alpha LM(\alpha, \beta_0)$ 5%	$\min_\alpha LM(\alpha, \beta_0)$ 10%
ϕ_x	7.6%	13.7%	2.1%	7.1%
ϕ_π	8.8%	15.8%	2.6%	7.1%
α	14.9%	27.1%	10.0%	19.5%
ρ	15.1%	27.5%	9.8%	19.8%
δ	12.6%	22.1%	5.5%	11.6%
κ	17.3%	27.3%	11.1%	19.7%
σ_a	15.2%	26.5%	10.4%	18.8%
σ_u	9.2%	16.2%	1.9%	7.3%
σ	16.4%	26.9%	10.1%	20.3%

parameters. Interestingly, when we consider the minimized statistic the tests we receive are quite conservative for some parameters while somewhat over-rejecting for others.

7.2 Nonlinear Weak IV

Figure 3: CDF of $\widetilde{LM}(\pi_0)$ for nonlinear weak IV, $C = .01$, T=100



In Section 5 we prove that, provided α is well identified, under appropriate assumptions $\widetilde{LM}(\beta_0)$ converges to a $\chi^2_{k_\beta}$ distribution asymptotically, where k_β is the dimension of β. As shown in Section 6, for the exponential family model where α is weakly identified but enters linearly we again have that $\widetilde{LM}(\beta_0)$ converges to a $\chi^2_{k_\beta}$. To understand the extent to which this result relies on the fact that α, the nuisance parameter, enters the expression linearly, we here consider a variation on the usual weak IV model in which β

enters the equation for Y nonlinearly. In particular, the model is:

$$Y = \pi\left(\beta^2 Z^2 + \beta Z\right) + U; \quad X = \pi Z + V$$

with β, π scalar and $(u_t, v_t)' \sim iid\, N(0, Id_2)$. As usual with weak IV, we take the first-stage parameter to zero as the sample size grows, $\pi = \frac{C}{\sqrt{T}}$. The log-likelihood for this model is $\ell(\theta) = const - \frac{1}{2}\sum (y_t - \pi(\beta^2 z_t^2 + \beta z_t))^2 - \frac{1}{2}\sum (x_t - \pi z_t)$. We consider testing $H_0 : \pi = \pi_0$ using $\widetilde{LM}(\pi_0)$ as defined in (5), and are interested in whether this statistic has a χ_1^2 distribution asymptotically. While we do not have any theoretical results for this case, we have run a number of simulations, which suggest that a χ_1^2 is a reasonable approximation to the distribution of this statistic. In particular, we set $\beta = 1$ and, $c = .01$, and consider $T = 100$ and $T = 10,000$. For each value of T, we simulate 10,000 Monte-Carlo draws, and calculate the size of asymptotic 5% and 10% tests (using critical values based on a χ_1^2) for sample sizes 100 and 10,000, which we report in Table 4. We also plot the CDF of $\widetilde{LM}(\pi_0)$, together with that of a χ_1^2, in Figure 3. These simulation results show that the distribution of $\widetilde{LM}(\pi_0)$ is close to a χ_1^2 in this model, suggesting that it may be possible to extend our theoretical results to this context.

Table 4: Size of 5% and 10% Tests based on $\widetilde{LM}(\pi_0)$for Nonlinear IV Model

Sample Size	Rejection rate for 5% test	Rejection rate for 10% test
100	6.49%	12.70%
10000	5.70%	11.34%

7.3 ARMA(1,1) with nearly canceling roots

Example 2 (cont.) We examine the performance of our proposed tests in the weak ARMA(1,1) model. We simulate samples of size 50 from the model

$$Y_t = (\pi + \beta)Y_{t-1} + e_t - \pi e_{t-1}, \quad e_t \sim i.i.d.N(0,1).$$

taking $\pi = .5$, $\beta = \frac{C}{\sqrt{T}}$ and $C = .01$. The simulated cdfs for $LM(\theta_0)$ and $\widetilde{LM}(\pi_0)$ are presented in figures 4 and 5, respectively. As these results make clear, the simulated distributions of both tests in this model are quite close to the their asymptotic distributions, even for moderate sample sizes.

Figure 4: CDF of $LM(\beta_0, \pi_0)$ for ARMA(1,1), $C = .01$, T=50

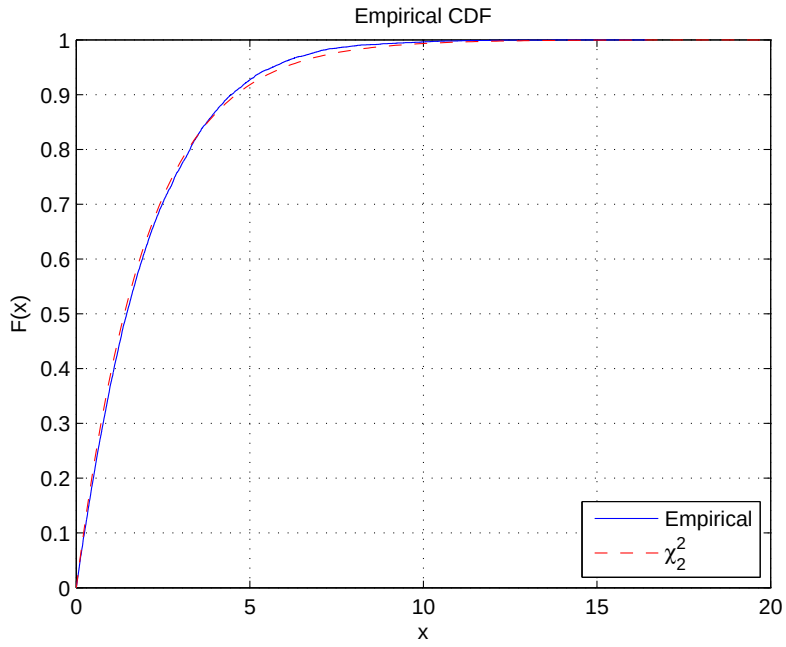
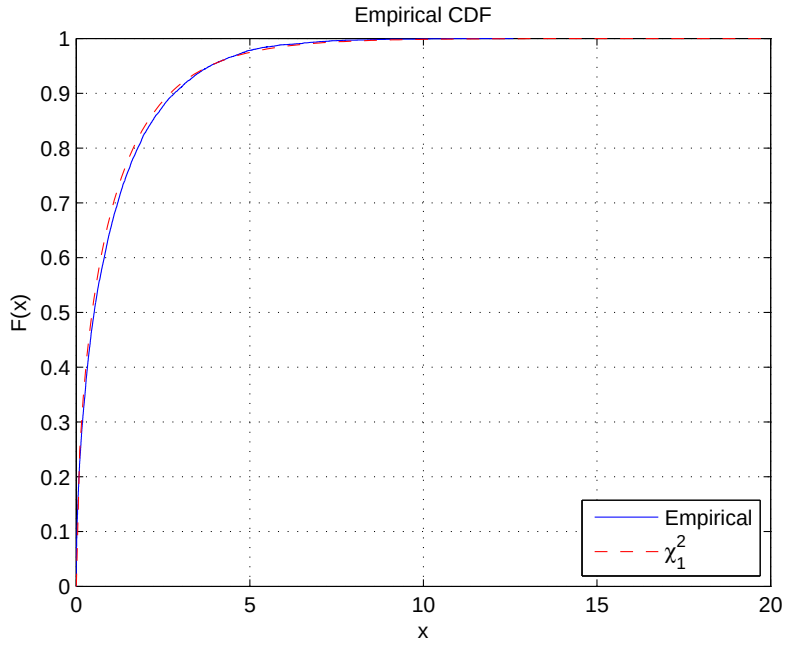


Figure 5: CDF of $\widetilde{LM}(\pi_0)$ for ARMA(1,1), $C = .01$, T=50



8 References

Andrews, D.W.K., and X. Cheng (2009): "Estimation and Inference with Weak, Semi-strong and Strong Identification," *unpublished manuscript*.

Andrews, D.W.K., M. Moreira, and J. Stock (2006): "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression," *Econometrica*, 74, 715-752.

Akritas, M. (1988): "An Asymptotic Derivation of Neyman's $C(\alpha)$ Test," *Statistics and Probablity Letters*, 6, 363-367.

Barndorff-Nielsen, O.E., and M. Sorensen (1991): "Information Quantities in Non-classical Settings," *Computational Statiatics and Data Analysis*, 12, 143-158.

Basawa, I.V., and H.L. Koul (1979): "Asymptotic Tests of Composite Hypotheses for Non-ergodic Type Stochastic Processes," *Stochastic Processes and their Applications*, 9, 291-305.

Bhat, B.R. (1974): "On the Method of Maximum-Likelihood for Dependent Observations," *Journal of the Royal Statistical Society. Series B (Methodological)*, 36, 48-53.

Canova, F., and L. Sala (2009): "Back to Square One: Identification Issues in DSGE Models," *Journal of Monetary Economics*, 56, 431-449.

Clarida, R., J. Gali, and M. Gertler (1999): "The Science of Monetary Policy: A New Keynesian Perspective," *Journal of Economic Literature*, 37, 1661-1707.

Crowder, M.J. (1976): "Maximum Likelihood Estimation for Dependent Observations," *Journal of the Royal Statistical Society. Series B (Methodological)*, 38, 45-53.

Dufor, J.M., L. Khalaf, and M. Kichian (2009): "Structural Multi-Equation Macroeconomic Models: Identification-Robust Estimation and Fit," *Bank of Canada Working Paper.*

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

Fernández-Villaverde, J. (2010): "The Econometrics of DSGE Models," *SERIES: Journal of the Spanish Economic Association*, 1, 3-49.

Fernández-Villaverde J., J. F. Rubio-Ramírez, T. J. Sargent, and M. W. Watson, (2007): "ABCs (and Ds) of Understanding VARs," *American Economic Review*, 97(3), 1021-1026.

Guerron-Quintana, P., A. Inoue, and L. Kilian (2009): "Frequentist Inference in Weakly Identified DSGE Models," *unpublished manuscript.*

Hall, P., and C.C. Heyde (1980): "Martingale Limit Theory and its Application," *Academic Press.*

Heijmans, R.D.H., and J.R. Magnus (1986): "Consistent Maximum-Likelihood Estimation with Dependent Observations," *Journal of Econometrics*, 32, 253-285.

Ingram, B. F. , N. R. Kocherlakota, and N. E. Savin (1994): "Explaining Business Cycles: A Multiple-shock Approach," *Journal of Monetary Economics* , 34, 415-428.

Ireland, P. N. (2004): "Technology Shocks in the New Keynesian Model," *The Review of Economics and Statistics*, 86, 923-936.

Iskrev, N. (2010): "Evaluating the Strength of Identification in DSGE Models. An a Priori Approach", *Bank of Portugal working paper*.

Jeganathan, P. (1995): "Some Aspect of Asymptotic Theory with Applications to Time Series Models," *Econometric Theory*, 11(5), 818-887.

Kleibergen, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781-1803.

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

Kleibergen, F., and S. Mavroeidis (2009): "Inference on Subsets of Parameters in GMM without Assuming Identification," *unpublished manuscript*.

Kocherlakota, S., and K. Kocherlakota (1991): "Neyman's $C(\alpha)$ Test and Rao's Efficient Score Test for Composite Hypotheses," *Statistics & Probability Letters*, 11, 491-493.

Komunjer, I., and S. Ng (2009): "Dynamic Identification of DSGE Models," *unpublished manuscript, University of California, San Diego, and Columbia University*.

Le Cam, L., and G.L. Yang (2000): "Asymptotics in Statistics: Some Basic Concepts," *Springer Series in Statistics*.

Lindé, J. (2005): "Estimating New-Keynesian Phillips Curves: A Full Information Maximum Likelihood Approach," *Journal of Monetary Economics*, 52(6), 1135-1149.

Liptser, R., and A. Shiryayev (1989): "Theory of Martingales," *Springer*.

McGrattan, E.R., R. Rogerson, and R. Wright (1997): "An Equilibrium Model of the Business Cycle with Household Production and Fiscal Policy," *International Economic Review*, 38(2), 267-290.

Moreira, M. (2001): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027-1048.

Schorfheide, F. (2010): "Estimation and Evaluation of DSGE Models: Progress and Challanges," *NBER Working Paper*.

Silvey, S.D. (1961): "A Note on Maximum-Likelihood in the Case of Dependent Random Variables," *Journal of the Royal Statistical Society. Series B* , 23, 444-452.

Staiger, D., and J.H. Stock (1997): "Instrumental Variables Regression With Weak Instruments," *Econometrica*, 65, 557-586.

Stock, J. H., and J.H. Wright (2000): "GMM With Weak Identification," *Econometrica*, 68, 1055-1096.

Stock, J.H., J.H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak

Identification in Generalized Method of Moments," *Journal of Business & Economic statistics*, 20(4), 518-529.

White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

White, H. (1982): "Maximum Likelihood Estimation in Misspecified Models," *Econometrica*, 50, 1-25.

9 Appendix with Proofs

We denote by super-script 0 quantities evaluated at $\theta_0 = (\alpha_0', \beta_0')'$. In the Taylor expansions used in the proof for Theorem 2, the expansion is assumed to be for each entry of the expanded matrix.

Proof of Lemma 1

The proof follows closely the argument of Bhat (1974), starting with the Taylor expansion:

$$0 = S_\alpha(\hat{\alpha}, \beta_0) = S_\alpha^0 - I_{\alpha\alpha}^0(\hat{\alpha} - \alpha_0) - (I_{\alpha\alpha}(\alpha^*, \beta_0) - I_{\alpha\alpha}^0)(\hat{\alpha} - \alpha_0),$$

where α^* is a convex combination of $\hat{\alpha}$ and α_0. We may consider different α^* for different rows of $I_{\alpha\alpha}$. Assumption 2(b) helps to control the last term of this expansion, while Assumption 2(a) allows us to substitute $J_{\alpha\alpha,T}$ for $I_{\alpha\alpha,T}$ in the second term. Assumption 1 gives the CLT for $K_{\alpha,T}S_{\alpha,T}$. □

Lemma 4 *Let $M_T = \sum_{t=1}^T m_t$ be a multi-dimensional martingale with respect to sigma-field $\mathcal{F}_t$, and let $[X]_t$ be its quadratic variation. Assume that there is a sequence of diagonal matrices K_T such that M_T satisfies the conditions of Assumption 3. Let $m_{i,t}$ be the i-th component of m_t, and $K_{i,T}$ the i-th diagonal element of K_T. For any i, j, l:*

$$K_{i,T}K_{j,T}K_{l,T}\sum_{t=1}^T m_{i,t}m_{j,t}m_{l,t} \to^p 0.$$

Proof of Lemma 4 Take any $\varepsilon > 0$,

$$\left|K_{i,T}K_{j,T}K_{l,T}\sum_{t=1}^T m_{i,t}m_{j,t}m_{l,t}\right| \leq \max_t |K_{i,T}m_{i,t}| \left|K_{j,T}K_{l,T}\sum_{t=1}^T m_{j,t}m_{l,t}\right| =$$

$$= \max_t |K_{i,T}m_{i,t}| \left|K_{j,T}K_{l,T}[M_j, M_l]_T\right|.$$

Assumption 3(b) implies that $K_{j,T}K_{l,T}[M_j, M_l]_T \to^p \Sigma_{j,l}$ is bounded in probability.

$$E\left(\max_t |K_{i,T}m_{i,t}|\right) \leq \varepsilon + E\left(K_{i,T}\max_t |m_{i,t}|\mathbb{I}\{|K_{i,T}m_{i,t}| > \varepsilon\}\right) \leq$$
$$\leq \varepsilon + \sum_t E\left(K_{i,T}|m_{i,t}|\mathbb{I}\{|K_{i,T}m_{i,t}| > \varepsilon\}\right).$$

The last term converges to 0 by Assumption 3(a). □

Proof of Lemma 2 Notice first that

$$-\frac{\partial}{\partial\alpha_i}I_{\alpha\alpha} = -[A_{\alpha\alpha_i}, S_\alpha] - [A_{\alpha\alpha}, S_{\alpha_i}] - [S_\alpha, A_{\alpha\alpha_i}] + 2\sum_{t=1}^T s_{\alpha,t}s'_{\alpha,t}s_{\alpha_i,t} + \Lambda_{\alpha\alpha\alpha_i}, \qquad (9)$$

where $\Lambda_{\alpha\alpha\alpha_i}$ is as defined in (7), and $[M, N] = \sum_{t=1}^T m_t n'_t$.

Denote by $f_t = f(x_t|X_{t-1};\theta)$ the (valid) pdf, while $f_{\alpha,t}$, $f_{\alpha\alpha,t}$ etc. are its partial derivatives with respect to α. Notice that the increments of $S_{\alpha,T}$, $A_{\alpha\alpha,T}$ and $\Lambda_{\alpha\alpha\alpha_i}$ are $s_{\alpha,t} = \frac{f_{\alpha,t}}{f_t}$, $a_{\alpha\alpha,t} = \frac{f_{\alpha\alpha,t}}{f_t}$, and $\lambda_{\alpha\alpha\alpha_i,t} = \frac{f_{\alpha\alpha\alpha_i,t}}{f_t}$ respectively. By definition

$$-\frac{\partial}{\partial\alpha_i}I_{\alpha\alpha} = \frac{\partial^3}{\partial\alpha\partial\alpha'\partial\alpha_i}\sum_t^T \log f_t = \sum_t^T \frac{f_{\alpha\alpha\alpha_i,t}}{f_t} - \sum_t^T \frac{f_{\alpha\alpha_i,t}}{f_t}\frac{f'_{\alpha,t}}{f_t} -$$
$$-\sum_t^T \frac{f_{\alpha\alpha,t}}{f_t}\frac{f_{\alpha_i,t}}{f_t} - \sum_t^T \frac{f_{\alpha,t}}{f_t}\frac{f'_{\alpha\alpha_i,t}}{f_t} + 2\sum_t^T \frac{f_{\alpha,t}}{f_t}\frac{f'_{\alpha,t}}{f_t}\frac{f_{\alpha_i,t}}{f_t},$$

so (9) follows.

Now consider the quantity of interest from Assumption (2b)

$$\left|K_{\alpha,T}(I_{\alpha\alpha}(\alpha_1,\beta_0) - I^0_{\alpha\alpha})K_{\alpha,T}\right| = \sum_i K_{\alpha,T}\left|\frac{\partial}{\partial\alpha_i}I^*_{\alpha\alpha}\right|K_{\alpha,T}|\alpha_{1,i} - \alpha_{0,i}|.$$

It suffices to show that $K_{\alpha_i,T}K_{\alpha,T}\left|\frac{\partial}{\partial\alpha_i}I^*_{\alpha\alpha}\right|K_{\alpha,T} \to^p 0$, using identity (9). Assumption (2b') implies that the last term converges to zero in probability. Lemma 4 implies that the second term is negligible. And finally, Assumption (2a) gives us that the first term also converges to zero in probability. □

Proof of Theorem 2 For simplicity of notation we assume in this proof that $C_{ij} = C$ for all i, j. The generalization of the proof to the case with different C_{ij}'s is obvious but

tedious. According to the martingale CLT, Assumption 3 implies that

$$(K_{\alpha,T}S^0_\alpha, K_{\beta,T}S^0_\beta, K_{\alpha\beta,T}vec(A^0_{\alpha\beta})') \Rightarrow (\xi_\alpha, \xi_\beta, \xi_{\alpha\beta}), \tag{10}$$

where the ξ's are jointly normal with variance matrix Σ_M.

We Taylor expand $S_{\beta_j}(\hat{\alpha}, \beta_0)$, the j-th component of vector $S_\beta(\hat{\alpha}, \beta_0)$, keeping in mind that $I^0_{\beta_j\alpha} = -\frac{\partial^2}{\partial\beta_j\partial\alpha}\ell(\alpha_0, \beta_0)$, and receive

$$K_{\beta_j,T}S_{\beta_j}(\hat{\alpha}, \beta_0) = K_{\beta_j,T}S^0_{\beta_j} - K_{\beta_j,T}I^0_{\beta_j\alpha}(\hat{\alpha} - \alpha_0) + \frac{1}{2}K_{\beta_j,T}(\hat{\alpha} - \alpha_0)'(I^0_{\alpha\alpha\beta_j})(\hat{\alpha} - \alpha_0) + \tilde{R}_j,$$

with residual

$$\tilde{R}_j = K_{\beta_j,T}\frac{1}{2}(\hat{\alpha} - \alpha_0)'(I^*_{\alpha\alpha\beta_j} - I^0_{\alpha\alpha\beta_j})(\hat{\alpha} - \alpha_0),$$

where $I^0_{\alpha\alpha\beta_j} = \frac{\partial^3}{\partial\alpha\partial\alpha'\partial\beta_j}\ell(\alpha_0, \beta_0)$, $I^*_{\alpha\alpha\beta_j} = \frac{\partial^3}{\partial\alpha\partial\alpha'\partial\beta_j}\ell(\alpha^*, \beta_0)$, and α^* is again a point between $\hat{\alpha}$ and α_0. From Lemma 1 we have that $K^{-1}_{\alpha,T}|\hat{\alpha} - \alpha_0| = O_p(1)$. As a result, Assumption 4 (c) makes the Taylor residual negligible:

$$K_{\beta_j,T}S_{\beta_j}(\hat{\alpha}, \beta_0) = K_{\beta_j,T}S^0_{\beta_j} - K_{\beta_j,T}I^0_{\beta_j\alpha}(\hat{\alpha} - \alpha_0) + \frac{1}{2}K_{\beta_j,T}(\hat{\alpha} - \alpha_0)'(I^0_{\alpha\alpha\beta_j})(\hat{\alpha} - \alpha_0) + o_p(1).$$

We plug asymptotic statement (6) into this equation and get

$$K_{\beta_j,T}S_{\beta_j}(\hat{\alpha}, \beta_0) = K_{\beta_j,T}S^0_{\beta_j} - K_{\beta_j,T}I^0_{\beta_j\alpha}(I^0_{\alpha\alpha})^{-1}S^0_\alpha + \frac{1}{2}K_{\beta_j,T}S^{0\prime}_\alpha(I^0_{\alpha\alpha})^{-1}(I^0_{\alpha\alpha\beta_j})(I^0_{\alpha\alpha})^{-1}S^0_\alpha + o_p(1).$$

Recall that by definition $I^0_{\beta\alpha} = J^0_{\beta\alpha} - A^0_{\beta\alpha}$. We use this substitution in the equation above, and receive:

$$\begin{aligned} K_{\beta_j,T}S_{\beta_j}(\hat{\alpha}, \beta_0) = K_{\beta_j,T}S^0_{\beta_j} - K_{\beta_j,T}J^0_{\beta_j\alpha}(I^0_{\alpha\alpha})^{-1}S^0_\alpha + K_{\beta_j,T}A^0_{\beta_j\alpha}(I^0_{\alpha\alpha})^{-1}S^0_\alpha + \\ + \frac{1}{2}K_{\beta_j,T}S^{0\prime}_\alpha(I^0_{\alpha\alpha})^{-1}(I^0_{\alpha\alpha\beta_j})(I^0_{\alpha\alpha})^{-1}S^0_\alpha + o_p(1). \end{aligned} \tag{11}$$

One can notice that we have the following informational equality:

$$I^0_{\alpha\alpha\beta_j} = -[A^0_{\alpha\alpha}, S^0_{\beta_j}] - [A^0_{\alpha\beta_j}, S^0_\alpha] - [S^0_\alpha, A^0_{\alpha\beta_j}] + 2\sum_{t=1}^{T} s_{\alpha,t}s'_{\alpha,t}s_{\beta_j,t} + \Lambda_{\alpha\alpha\beta_j}. \tag{12}$$

It can be obtained in the same manner as (9). Assumption 4(b) implies that $K_{\beta_j,T}K_{\alpha,T}\Lambda_{\alpha\alpha\beta_j}K_{\alpha,T} \to^p 0$. Assumption 2(a) and Assumption 3 together imply that $(K_{\alpha,T} \otimes K_{\alpha,T})K^{-1}_{\alpha\alpha,T} \to 0$. Using Assumption 2(a) and Lemma 4, we notice that

$$K_{\beta_j,T}K_{\alpha,T}I^0_{\alpha\alpha\beta_j}K_{\alpha,T} = -K_{\beta_j,T}K_{\alpha,T}[A^0_{\alpha\beta_j}, S^0_\alpha]K_{\alpha,T} - K_{\beta_j,T}K_{\alpha,T}[S^0_\alpha, A^0_{\alpha\beta_j}]K_{\alpha,T} + o_p(1). \tag{13}$$

According to Assumption 4(a), $K_{\beta_j,T}K_{\alpha,T}[A^0_{\alpha\beta_j}, S^0_\alpha]K_{\alpha,T}$ is asymptotically bounded so

$K_{\beta_j,T}K_{\alpha,T}I^0_{\alpha\alpha\beta_j}K_{\alpha,T} = O_p(1)$. By Assumption 2(a) $K_{\alpha,T}I^0_{\alpha\alpha}K_{\alpha,T} = K_{\alpha,T}J_{\alpha\alpha}K_{\alpha,T} + o_p(1)$; Assumption 4(a) implies that $K_{\alpha,T}A_{\alpha\beta}K_{\beta,T}$ is bounded. Taken together, these statements imply that we can substitute $J^0_{\alpha\alpha}$ for $I^0_{\alpha\alpha}$ everywhere in (11). Doing so gives us:

$$\begin{aligned} K_{\beta_j,T}S_{\beta_j}(\hat{\alpha},\beta_0) = K_{\beta_j,T}S^0_{\beta_j} - K_{\beta_j,T}J^0_{\beta_j\alpha}(J^0_{\alpha\alpha})^{-1}S^0_\alpha + K_{\beta_j,T}A^0_{\beta_j\alpha}(J^0_{\alpha\alpha})^{-1}S^0_\alpha + \\ + \frac{1}{2}K_{\beta_j,T}S^{0'}_\alpha(J^0_{\alpha\alpha})^{-1}(I^0_{\alpha\alpha\beta_j})(J^0_{\alpha\alpha})^{-1}S^0_\alpha + o_p(1), \\ K_{\beta_j,T}S_{\beta_j}(\hat{\alpha},\beta_0) = K_{\beta_j,T}S^0_{\beta_j} - K_{\beta_j,T}J^0_{\beta_j\alpha}(J^0_{\alpha\alpha})^{-1}S^0_\alpha + D'_j(J^0_{\alpha\alpha}K_{\alpha,T})^{-1}S^0_\alpha + o_p(1), \quad (14) \end{aligned}$$

where

$$D_j = K_{\alpha,T}K_{\beta_j,T}A^0_{\alpha\beta_j} + \frac{1}{2}K_{\alpha,T}K_{\beta_j,T}(I^0_{\alpha\alpha\beta_j})(J^0_{\alpha\alpha})^{-1}S^{0'}_\alpha.$$

Notice that D, a $k_\alpha \times k_\beta$ random matrix, is asymptotically normal (though it may have zero variance, i.e. it may converge to zero) and asymptotically independent of $K_{\alpha,T}S^0_\alpha$. Indeed, using (13) we have:

$$\begin{aligned} D_j =& K_{\alpha,T}K_{\beta_j,T}K^{-1}_{\alpha\beta_j,T}\left(K_{\alpha\beta_j,T}A^0_{\alpha\beta_j} - (K_{\alpha\beta_j,T}[A^0_{\alpha\beta_j}, S^0_\alpha]K_{\alpha,T})(K_{\alpha,T}J^0_{\alpha\alpha}K_{\alpha,T})^{-1}K_{\alpha,T}S^{0'}_\alpha\right) + o_p(1) \Rightarrow \\ \Rightarrow& C\left(\xi_{\alpha\beta_j} - cov(\xi_{\alpha\beta_j},\xi_\alpha)Var(\xi_\alpha)^{-1}\xi_\alpha\right), \end{aligned}$$

where variables $(\xi'_\alpha, \xi'_{\alpha\beta_j}) = \lim(K_{\alpha,T}S^{0'}_\alpha, K_{\alpha\beta_j,T}A^{0'}_{\alpha\beta_j})$ are as described at the beginning of the proof.

Plugging the last statement and (10) into equation (14) we have:

$$\begin{aligned} K_{\beta_j,T}S_{\beta_j}(\hat{\alpha},\beta_0) \Rightarrow \xi_{\beta_j} - cov(\xi_{\beta_j},\xi_\alpha)Var(\xi_\alpha)^{-1}\xi_\alpha + \\ + C\left(\xi_{\alpha\beta_j} - cov(\xi_{\alpha\beta_j},\xi_\alpha)Var(\xi_\alpha)^{-1}\xi_\alpha\right)Var(\xi_\alpha)^{-1}\xi_\alpha. \quad (15) \end{aligned}$$

Conditional on ξ_α, $K_{\beta,T}S_\beta(\hat{\alpha},\beta_0)$ is an asymptotically normal vector with mean zero.

Now we turn to the inverse variance term in formula (5) for $\widetilde{LM}(\beta_0)$, $\left(J_{\beta\beta} - J_{\beta\alpha}J^{-1}_{\alpha\alpha}J'_{\beta\alpha}\right)\big|_{(\hat{\alpha},\beta_0)}$. Below we prove the following lemma:

Lemma 5 *Under the Assumptions of Theorem 2 we have:*

(a) $K_{\beta_i,T}K_{\beta_j,T}J_{\beta_i\beta_j}(\hat{\alpha},\beta_0) \Rightarrow cov(\xi_{\beta_i},\xi_{\beta_j}) + C\cdot cov(\xi_{\alpha\beta_i},\xi_{\beta_j})'Var(\xi_\alpha)^{-1}\xi_\alpha +$
$+ C\cdot cov(\xi_{\alpha\beta_j},\xi_{\beta_i})'Var(\xi_\alpha)^{-1}\xi_\alpha + C^2\xi'_\alpha Var(\xi_\alpha)^{-1}cov(\xi_{\alpha\beta_i},\xi_{\alpha\beta_j})Var(\xi_\alpha)^{-1}\xi_\alpha;$

(b) $K_{\alpha,T}K_{\beta_j,T}J_{\alpha\beta_j}(\hat{\alpha},\beta_0) \Rightarrow cov(\xi_\alpha,\xi_{\beta_j}) + C\cdot cov(\xi_{\alpha\beta_j},\xi_\alpha)Var(\xi_\alpha)^{-1}\xi_\alpha;$

(c) $K_{\alpha,T}J_{\alpha\alpha}(\hat{\alpha},\beta_0)K_{\alpha,T} \to^p Var(\xi_\alpha).$

Lemma 5 implies that

$$
\begin{aligned}
&K_{\beta_i,T}K_{\beta_j,T}\left(J_{\beta_i\beta_j}-J_{\beta_i\alpha}J_{\alpha\alpha}^{-1}J'_{\beta_j\alpha}\right)\Big|_{(\hat{\alpha},\beta_0)}\Rightarrow\\
&\Rightarrow cov(\xi_{\beta_i},\xi_{\beta_j})+C\cdot cov(\xi_{\alpha\beta_i},\xi_{\beta_j})'Var(\xi_\alpha)^{-1}\xi_\alpha+C\cdot cov(\xi_{\alpha\beta_j},\xi_{\beta_i})'Var(\xi_\alpha)^{-1}\xi_\alpha\\
&+C^2\xi'_\alpha Var(\xi_\alpha)^{-1}cov(\xi_{\alpha\beta_i},\xi_{\alpha\beta_j})Var(\xi_\alpha)^{-1}\xi_\alpha-\\
&-\left(cov(\xi_\alpha,\xi_{\beta_i})+C\cdot cov(\xi_{\alpha\beta_i},\xi_\alpha)Var(\xi_\alpha)^{-1}\xi_\alpha\right)'Var(\xi_\alpha)^{-1}\times\\
&\times\left(cov(\xi_\alpha,\xi_{\beta_j})+C\cdot cov(\xi_{\alpha\beta_j},\xi_\alpha)Var(\xi_\alpha)^{-1}\xi_\alpha\right).
\end{aligned}
$$

Note that the last expression is the same as the variance the right side of equation (15) conditional on random variable ξ_α. That is, $K_{\beta,T}\left(J_{\beta\beta}-J_{\beta\alpha}J_{\alpha\alpha}^{-1}J'_{\beta\alpha}\right)K_{\beta,T}\big|_{(\hat{\alpha},\beta_0)}$ is asymptotically equal to the asymptotic variance of $K_{\beta,T}S_\beta(\hat{\alpha},\beta_0)$ conditional on ξ_α. As a result statistic $\widetilde{LM}(\beta_0)$, conditional on ξ_α, is distributed $\chi^2_{k_\beta}$ asymptotically and thus is asymptotically $\chi^2_{k_\beta}$ unconditionally as well. This completes the proof of Theorem 2.

Proof of Lemma 5

(a) We can Taylor expand $J_{\beta_i\beta_j}(\hat{\alpha},\beta_0)$ as:

$$
J_{\beta_i\beta_j}(\hat{\alpha},\beta_0)=J^0_{\beta_i\beta_j}+\frac{\partial}{\partial\alpha}J^0_{\beta_i\beta_j}(\hat{\alpha}-\alpha_0)+\frac{1}{2}(\hat{\alpha}-\alpha_0)'\frac{\partial^2}{\partial\alpha\partial\alpha'}J^0_{\beta_i\beta_j}(\hat{\alpha}-\alpha_0)+R_{ij},\qquad(16)
$$

where

$$
K_{\beta_i,T}K_{\beta_j,T}R_{ij}=K_{\beta_i,T}K_{\beta_j,T}\frac{1}{2}(\hat{\alpha}-\alpha_0)'\left(\frac{\partial^2}{\partial\alpha\partial\alpha'}J^0_{\beta_i\beta_j}-\frac{\partial^2}{\partial\alpha\partial\alpha'}J^*_{\beta_i\beta_j}\right)(\hat{\alpha}-\alpha_0)
$$

is negligible asymptotically due to Assumption 4(c). Consider the first term of the Taylor expansion above:

$$
\frac{\partial}{\partial\alpha}J_{\beta_i\beta_j}=\frac{\partial}{\partial\alpha}\sum_t s_{\beta_i,t}s_{\beta_j,t}=[A_{\alpha,\beta_i},S_{\beta_j}]+[A_{\alpha,\beta_j},S_{\beta_i}]-2\sum s_{\alpha,t}s_{\beta_i,t}s_{\beta_j,t}.
$$

Using Lemma 4 and Assumption 4(a) we have

$$
K_{\alpha,T}K_{\beta_i,T}K_{\beta_j,T}\frac{\partial}{\partial\alpha'}J_{\beta_i\beta_j}\to^p C\cdot cov(\xi_{\alpha\beta_i},\xi_{\beta_j})+C\cdot cov(\xi_{\alpha\beta_j},\xi_{\beta_i}).\qquad(17)
$$

Now let us consider the normalized second derivative of $J_{\beta_i\beta_j}$:

$$K_{\beta_i,T}K_{\beta_j,T}K_{\alpha,T}\frac{\partial^2}{\partial\alpha\partial\alpha'}J_{\beta_i\beta_j}K_{\alpha,T} =$$
$$= K_{\beta_i,T}K_{\beta_j,T}K_{\alpha,T}\left([\Lambda_{\alpha\alpha\beta_i},S_{\beta_j}]+[\Lambda_{\alpha\alpha\beta_j},S_{\beta_i}]+[A_{\alpha\beta_i},A_{\alpha\beta_j}]+[A_{\alpha\beta_j},A_{\alpha\beta_i}]\right)K_{\alpha,T}+o_p(1).$$

The $o_p(1)$ term appears due to Lemma 4, applied to the remaining terms. Assumption 4(b) implies that $K_{\alpha,T}K_{\beta_i,T}K_{\beta_j,T}[\Lambda_{\alpha\alpha\beta_i},S_{\beta_j}]K_{\alpha,T}\to^p 0$. Finally using Assumption 3(b) we get

$$K_{\beta_i,T}K_{\beta_j,T}K_{\alpha,T}\frac{\partial^2}{\partial\alpha\partial\alpha'}J_{\beta_i\beta_j}K_{\alpha,T}\to^p C^2cov(\xi_{\alpha\beta_i},\xi_{\alpha\beta_j})+C^2cov(\xi_{\alpha\beta_j},\xi_{\alpha\beta_i}). \tag{18}$$

Putting the expressions for derivatives (17) and (18) into equation (16), and also noticing that due to Lemma 1 $K_{\alpha,T}^{-1}(\hat{\alpha}-\alpha_0)\Rightarrow Var(\xi_\alpha)^{-1}\xi_\alpha$, we get statement (a).

(b) Again we use Taylor expansion:

$$J_{\alpha\beta_j}(\hat{\alpha},\beta_0)=J^0_{\alpha\beta_j}+\frac{\partial}{\partial\alpha}J^0_{\alpha\beta_j}(\hat{\alpha}-\alpha_0)+\frac{1}{2}\sum_n\frac{\partial^2}{\partial\alpha\partial\alpha_n}J^*_{\alpha\beta_j}(\hat{\alpha}-\alpha_0)(\hat{\alpha}_n-\alpha_{0,n}). \tag{19}$$

From assumption 3(b)

$$K_{\alpha,T}K_{\beta_j,T}J^0_{\alpha\beta_j}\to^p cov(\xi_\alpha,\xi_{\beta_j}). \tag{20}$$

Taking the derivative we see

$$\frac{\partial}{\partial\alpha}J_{\alpha\beta_j}=\frac{\partial}{\partial\alpha}\sum_t s_{\alpha,t}s_{\beta_j,t}=[A_{\alpha\alpha},S_{\beta_j}]+[S_\alpha,A_{\alpha\beta_j}]-2\sum s_{\alpha,t}s'_{\alpha,t}s_{\beta_j,t}.$$

According to Lemma 4 $K_{\alpha,T}K_{\beta_j,T}\sum s_{\alpha,t}s'_{\alpha,t}s_{\beta_j,t}K_{\alpha,T}\to 0$. Assumptions 2(a) and 3 imply that $K_{\alpha,T}K_{\beta_j,T}[A_{\alpha\alpha},S_{\beta_j}]K_{\alpha,T}\to^p 0$. We have

$$K_{\alpha,T}K_{\beta_j,T}\frac{\partial}{\partial\alpha}J_{\alpha\beta_j}K_{\alpha,T}=K_{\alpha,T}K_{\beta_j,T}[S_\alpha,A_{\alpha\beta_j}]K_{\alpha,T}+o_p(1)\to^p C\cdot cov(\xi_\alpha,\xi_{\alpha\beta_j}).$$

Similarly, we can show that the residual term in (19) is asymptotically negligible. Putting the last equation, together with (20), into (19) and using Lemma 1 we get statement (b) of Lemma 5.

(c) As before we use Taylor expansion

$$K_{\alpha,T}J_{\alpha\alpha}(\hat{\alpha},\beta_0)K_{\alpha,T} = K_{\alpha,T}J^0_{\alpha\alpha}K_{\alpha,T} + \sum_n K_{\alpha,T}\frac{\partial}{\partial\alpha_n}J^*_{\alpha\alpha}(\hat{\alpha}_n - \alpha_{0,n})K_{\alpha,T};$$

$$\frac{\partial}{\partial\alpha_n}J_{\alpha\alpha} = [A_{\alpha\alpha_n}, S_\alpha] + [S_\alpha, A_{\alpha\alpha_n}] + 2\sum s_{\alpha,t}s'_{\alpha,t}s_{\alpha_n,t}.$$

By the same argument as before $K_{\alpha,T}K_{\alpha_n,T}[A_{\alpha\alpha_n}, S_\alpha]K_{\alpha,T} \to^p 0$, and according to Lemma 4, $K_{\alpha,T}K_{\alpha_n,T}\sum s_{\alpha,t}s'_{\alpha,t}s_{\alpha_n,t}K_{\alpha,T} \to^p 0$. Given the result of Lemma 1 we arrive at statement (c). □

Proof of Theorem 3. Whenever a function is given with no argument, it means it is evaluated at the true θ_0. For the functions ℓ, m, n and r only, subscript 1 stands for the partial derivative with respect to α_1, subscript 2 for the partial derivative with respect to α_2, and subscript β for the partial derivative with respect to β. M' denotes the transpose of M. For simplicity of notation this proof assumes that α_1 and α_2 are scalars. The generalization to the multidimensional case is obvious but tedious.

Let $H_1 = \sum_{t=1}^T \left(H(x_t) - \dot{A}\right)$ be a $p\times 1$ vector, $H_2 = \sum_{t=1}^T \left(H(x_t) - \dot{A}\right)\left(H(x_t) - \dot{A}\right)'$ be a $p\times p$ matrix. According to the conditions of Theorem 3, $\frac{1}{T}H_1 \to^p 0$ and $\frac{1}{T}H_2 \to^p -\ddot{A}$, and a Central Limit Theorem holds for $\frac{1}{\sqrt{T}}H_1$.

Consider the following normalization: $K_{\alpha_1,T} = \frac{1}{\sqrt{T}}; K_{\alpha_2,T} = 1; K_{\beta,T} = Id_{k_\beta}$. Below we check Assumptions 1, 2, 3 and 4 for the exponential model.

Assumption 1 One can check that

$$S_T = \begin{pmatrix} (m_1 + \frac{1}{\sqrt{T}}n_1\alpha_2 + \frac{1}{\sqrt{T}}r_1)'H_1 \\ \frac{n'}{\sqrt{T}}H_1 \\ \frac{(n_\beta\alpha_2 + r_\beta)'}{\sqrt{T}}H_1 \end{pmatrix} = \frac{\partial\eta}{\partial\theta}'H_1,$$

where $\frac{\partial\eta}{\partial\theta} = ((m_1 + \frac{1}{\sqrt{T}}n_1\alpha_2 + \frac{1}{\sqrt{T}}r_1), \frac{n}{\sqrt{T}}, \frac{(n_\beta\alpha_2+r_\beta)}{\sqrt{T}})$ is $p\times k$ matrix, $k = \dim(\beta) + 2$. It is easy to show that $J_T = \frac{\partial\eta}{\partial\theta}'H_2\frac{\partial\eta}{\partial\theta}$. Using the normalization K_T we have:

$$K_TJ_TK_T \to^p -\begin{pmatrix} m'_1 \\ n' \\ (n_\beta\alpha_2 + r_\beta)' \end{pmatrix}\ddot{A}(m_1, n, n_\beta\alpha_2 + r_\beta) = \Sigma.$$

Due to the rank assumption, Σ is positive-definite.

Assumption 2(a) We calculate $I_{\alpha\alpha,T}$:

$$I_{\alpha\alpha,T} = \begin{pmatrix} m_{11}'H_1 - Tm_1'\ddot{A}m_1 & \frac{n_1'}{\sqrt{T}}H_1 - T\frac{n'}{\sqrt{T}}\ddot{A}m_1 \\ \frac{n_1'}{\sqrt{T}}H_1 - T\frac{n'}{\sqrt{T}}\ddot{A}m_1 & -T\frac{n'}{\sqrt{T}}\ddot{A}\frac{n}{\sqrt{T}} \end{pmatrix} + o_p(1).$$

Now it is straightforward to show that $K_{\alpha,T}I_{\alpha\alpha,T}K_{\alpha,T}$ converges to the same limit as $K_{\alpha,T}J_{\alpha\alpha,T}K_{\alpha,T}$. This means that $J_{\alpha\alpha,T}^{-1}I_{\alpha\alpha,T} \to^p Id_2$.

Assumption 2(b) We can prove by tedious differentiation that

$$K_{\alpha_i,T}K_{\alpha_j,T}K_{\alpha_l,T}\frac{\partial^3\ell}{\partial\alpha_i\partial\alpha_j\partial\alpha_k} \to^p 0. \tag{21}$$

Below we drop all terms that are of obviously smaller order:

$$\frac{1}{T^{3/2}}\ell_{111} = \frac{1}{T^{3/2}}\left(H_1'm_{111} - 3Tm_1'\ddot{A}m_{11} - T\sum_i m_1'\dddot{A}_i m_1(m_1)_i\right) + o_p(1) \to^p 0;$$

$$\frac{1}{T}\ell_{112} = \frac{1}{T}\left(H_1'\frac{n_{11}}{\sqrt{T}} - Tm_{11}'\ddot{A}\frac{n}{\sqrt{T}} - 2Tm_1'\ddot{A}\frac{n_1}{\sqrt{T}} - T\sum_i m_1'\dddot{A}_i\frac{n}{\sqrt{T}}(m_1)_i\right) \to^p 0;$$

$$\frac{1}{\sqrt{T}}\ell_{122} = \frac{1}{\sqrt{T}}\left(T\sum_i \frac{n'}{\sqrt{T}}\dddot{A}_i\frac{n}{\sqrt{T}}(m_1)_i - 2T\frac{n'}{\sqrt{T}}\ddot{A}\frac{n_1}{\sqrt{T}}\right) \to 0;$$

$$\ell_{222} = -T\sum_i \frac{n}{\sqrt{T}}\dddot{A}_i\frac{n}{\sqrt{T}}\frac{(n)_i}{\sqrt{T}} \to 0,$$

here $\dddot{A}_i = \frac{\partial}{\partial\eta_i}\ddot{A}$, $(x)_i$ is the i-th component of vector x, and the summation runs over all components of η in the term involving the third derivative. The last two statements employ that α_2 enters linearly, so any time we differentiate with respect to α_2 a term including $\frac{n}{\sqrt{T}}$ appears. Given the third informational equality stated in (9) and Lemma 4, equation (21) implies that Assumption 2(b) holds.

Assumption 3 From the definition of $A_{\alpha\beta}$ one can see that:

$$\frac{A_{\beta\alpha_1}}{\sqrt{T}} = \frac{1}{\sqrt{T}}\frac{n_{\beta 1}'}{\sqrt{T}}H_1 + o_p(1), \quad A_{\beta\alpha_2} = \frac{n_\beta'}{\sqrt{T}}H_1 + o_p(1).$$

From the assumptions of Theorem 3 we get that $\frac{A_{\beta\alpha_1}}{\sqrt{T}} \to^p 0$ and that $A_{\beta\alpha_2}$ satisfies the Central Limit Theorem jointly with $S_T(\theta_0)$, where we use $K_{\beta\alpha_1} = \frac{1}{\sqrt{T}}$ and $K_{\beta\alpha_2} = 1$.

Assumption 4 Assumption 4(a) holds trivially. For Assumption 4(b) we check

that

$$\frac{1}{T}\ell_{11\beta} \to^p 0, \tag{22}$$

$$\frac{1}{\sqrt{T}}\left(\ell_{\beta 1\alpha_2} - [A_{\beta\alpha_2}, S_{\alpha_1}]\right) \to^p 0, \tag{23}$$

$$\ell_{22\beta} - 2[A_{\beta\alpha_2}, S_{\alpha_2}] \to^p 0. \tag{24}$$

Equation (22) comes from differentiation:

$$\frac{1}{T}\ell_{11\beta} = \frac{1}{T}\left(\left(\frac{n_{11\beta}\alpha_2 + r_{11\beta}}{\sqrt{T}}\right)' H_1 - T\left(\frac{n_\beta\alpha_2 + r_\beta}{\sqrt{T}}\right)' \ddot{A}m_{11} - \right.$$
$$\left. -2T\left(\frac{n_{1\beta}\alpha_2 + r_{1\beta}}{\sqrt{T}}\right)' \ddot{A}m_1 - T\sum_i \left(\frac{n_\beta\alpha_2 + r_\beta}{\sqrt{T}}\right)' \dddot{A}_i m_1 (m_1)_i\right) \to^p 0.$$

Taking derivatives one can check that $\frac{1}{\sqrt{T}}\ell_{12\beta} \to^p -n'_\beta \ddot{A}m_1$ and $A_{\beta\alpha_2} = \left(\frac{n_\beta}{\sqrt{T}}\right)' H_1 + o_p(1)$. As a result,

$$[A_{\beta\alpha_2}, \frac{S_{\alpha_1}}{\sqrt{T}}] = \frac{1}{T}n'_\beta H_2 m_1 \to^p -n'_\beta \ddot{A}m_1,$$

and statement (23) holds. We also have

$$[A_{\beta\alpha_2}, S_{\alpha_2}] = \frac{n'_{\beta_j}}{\sqrt{T}}H_2\frac{n}{\sqrt{T}} + o_p(1) \to^p -n'_{\beta_j}\ddot{A}n.$$

One can easily check that $\ell_{22} = -n'\ddot{A}n$, so we have $\ell_{22\beta} \to^p -2n'_\beta \ddot{A}n$. Together, these results imply (24). According to the third informational equality (a version of which is given in (12)) and Lemma 4 statements (22), (23) and (24) imply Assumption 4(b).

Assumption 4(c) It is enough to check that

$$\frac{1}{T^{3/2}}\ell_{111\beta} \to^p 0; \frac{1}{T}\ell_{112\beta} \to^p 0; \frac{1}{\sqrt{T}}\ell_{122\beta} \to^p 0; \ell_{222\beta} \to^p 0.$$

The idea here is that since η_T is linear in α_2, each additional derivative with respect to α_2 generates $\frac{n}{\sqrt{T}}$. If the derivative is taken with respect to α_1, then the additional normalization $1/\sqrt{T}$ is added. In any case the normalization of all terms will be excessively strong, so they will be asymptotically negligible.

We have shown that Assumptions 1-4 hold for the exponential model described in Theorem 3. Thus, applying Theorem 2, the conclusion of Theorem 3 holds. □