Quantitative Economics 6 (2015), 123–152 1759-7331/20150123

Maximum likelihood inference in weakly identified dynamic stochastic general equilibrium models

Isaiah Andrews
Department of Economics, MIT

Anna Mikusheva
Department of Economics, MIT

This paper examines the issue of weak identification in maximum likelihood, motivated by problems with estimation and inference in a multidimensional dynamic stochastic general equilibrium model. We show that two forms of the classical score (Lagrange multiplier) test for a simple hypothesis concerning the full parameter vector are robust to weak identification. We also suggest a test for a composite hypothesis regarding a subvector of parameters. The suggested subset test is shown to be asymptotically exact when the nuisance parameter is strongly identified. We pay particular attention to the question of how to estimate Fisher information and we make extensive use of martingale theory.

Keywords. Maximum likelihood, $C(\alpha)$ test, score test, weak identification.

JEL classification. C32.

1. Introduction

In recent years, we have witnessed the rapid growth of the empirical literature on the highly parameterized, microfounded 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, Altug (1989), 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

Isaiah Andrews: iandrews@mit.edu
Anna Mikusheva: amikushe@mit.edu

We would like to thank Patrik Guggenberger, Ulrich Muller, Whitney Newey, Serena Ng, Zhongjun Qu, Frank Schorfheide, Jim Stock, the anonymous referees, and seminar participants at the Winter Econometric Society Meeting in Chicago, Boston College, Canadian Econometric Study Group, Columbia, Harvard–MIT, NBER summer institute, Rice, Texas A&M, UC Berkeley, UC San Diego, UPenn, U Virginia, and Yale for helpful comments. We are grateful to Lynda Khalaf for guidance on implementing the procedure of Dufour, Khalaf, and Kichian (2013). Andrews gratefully acknowledges support from the Ford Foundation and the NSF Graduate Research Fellowship under Grant 1122374. Mikusheva gratefully acknowledges financial support from the Castle–Krob Career Development Chair and the Sloan Research Fellowship.

DOI: 10.3982/QE331

nearly flat surfaces...the standard errors of the estimates are notoriously difficult to compute and their asymptotic distribution a poor approximation to the small sample one." The poor performance of maximum likelihood estimation has fueled growing concerns about weak identification in many DSGE models (see Canova and Sala (2009), Guerron-Quintana, Inoue, and Kilian (2013), Iskrev (2010), and Mavroeidis (2005)).

In this paper, we consider the problem of weak identification in models estimated by maximum likelihood, focusing in particular on weakly identified DSGE models. 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, but 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 provide several examples illustrating ways in which weak identification may arise in a DSGE context.[1]

We focus on the problem of testing and confidence set construction in this context. We consider two different tasks. First, we examine the problem of testing a simple hypothesis on the full parameter vector. We suggest using particular forms of the classical Lagrange multiplier (LM) test, which we show are robust to weak identification. The assumptions needed for this result are extremely weak and cover a large number of cases, including all of our examples. An advantage of our approach is that we can remain agnostic about the source and nature of weak identification, and need not rely on any particular asymptotic embedding. The proof for these tests 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.

Second, we turn to the problem of testing a subset of parameters without restricting the remaining parameters. The tests we suggest for a subset of parameters are particular forms of Rao's score test and are asymptotically equivalent to Neyman's $C(\alpha)$ test when identification is strong. Consequently, our tests are efficient when all parameters are strongly identified. We show that the suggested tests have a χ^2 asymptotic distribution as long as the nuisance parameter (i.e., the part of the parameter vector that we are not testing) is strongly identified, even when the tested parameter is weakly identified. By combining our procedure for concentrating out nuisance parameters that are strongly identified with projection over the remaining nuisance parameters, one obtains weak identification-robust tests more powerful than those based on projection alone.

The paper also reveals a previously unnoticed fact concerning estimation of the Fisher information. White (1982) noted that in strongly identified models, the Fisher information can be estimated using either the Hessian of the likelihood or the quadratic variation of the score, and argued that a large discrepancy between these two estimates indicates model misspecification. We show in examples that weak identification leads to a distinct but related phenomenon. In particular, under weak identification, the appropriately normalized quadratic variation of the score converges to fixed positive-definite

[1] Due to space limitations, most of the examples are placed in a Supplement, available as a supplementary file on the journal website, http://qeconomics.org/supp/331/supplement.pdf.

matrix while the Hessian converges in distribution to a random matrix. Thus, large disparities between different estimators of information may arise even in correctly specified models if identification is weak.

The issue of weak identification in DSGE models was first highlighted by Mavroeidis (2005) and Canova and Sala (2009), who pointed out that the objective functions implied by many DSGE models are nearly flat in some directions. Weak identification-robust inference procedures for log-linearized DSGE models were introduced by Dufour, Khalaf, and Kichian (2013; henceforth DKK), Guerron-Quintana, Inoue, and Killian (2013; henceforth GQIK), and Qu (forthcoming). With the exception of GQIK, these papers focus on tests for the full parameter vector and make extensive use of the projection method to construct confidence sets for subsets of the structural parameters that, 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.

The LM tests we suggest in this paper can be applied whenever the correct likelihood is specified and, in particular, can accommodate nonlinear DSGE models, which are increasingly popular and cannot be treated by existing weak identification-robust methods. We compare our LM tests with the existing weak identification-robust methods from a theoretical perspective, and report an extensive simulation study in a small-scale DSGE model, demonstrating the advantages and disadvantages of different robust methods. In simulation, we find that our LM statistics have much higher power than the limited information tests suggested by DKK. The test statistic proposed by Qu (forthcoming) is almost indistinguishable from our LM_e statistic, but is defined for a much more limited set of models. The test of GQIK has power comparable to the LM tests in our simulation example, but is highly computationally intensive and relies on the questionable assumption of strong identification of the reduced-form parameters. Furthermore, this test will typically be asymptotically inefficient under strong identification of the structural parameters.

Structure of the paper. In Section 2, we discuss how weak identification can arise in DSGE models. Section 3 introduces our notation as well as some results from martingale theory; it also discusses the difference between two alternative measures of information. Section 4 suggests a test for the full parameter vector. Section 5 suggests a test for a hypothesis about a subset of parameters under the assumption that the nuisance parameter is strongly identified. Section 6 contains suggestions for applied researchers. Simulations supporting our theoretical results and comparing our procedures to existing alternatives are reported in Section 7. Section 8 concludes. Proofs of secondary importance, additional derivations, and further examples can be found in the Supplement. Replication files are also available on the journal website, http://qeconomics.org/supp/331/code_and_data.zip.

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 $[\cdot,\cdot]$ stands for the joint quadratic variation of two martingales; $\Rightarrow$ denotes weak convergence (convergence in distribution), while $\xrightarrow{p}$ stands for convergence in probability.

2. Weak identification in DSGE models

We begin by considering a highly stylized DSGE model that is much simpler than contemporary models designed to fit the data. Unlike most DSGE models used in empirical practice, this model can be solved analytically and allows us to demonstrate how weak identification can arise in a DSGE context.

Assume we observe data on inflation π_t and a measure of real activity x_t for periods $t = 1, \dots, T$. Assume that the dynamics of the data are described by the simple DSGE model

$$\begin{aligned} &bE_t\pi_{t+1} + \kappa x_t - \pi_t + \varepsilon_t = 0, \\ &-[r_t - E_t\pi_{t+1} - \rho\Delta a_t] + E_t x_{t+1} - x_t = 0, \\ &\lambda r_{t-1} + (1-\lambda)\phi_\pi \pi_t + (1-\lambda)\phi_x x_t + u_t = r_t. \end{aligned} \tag{1}$$

The first equation is a Phillips curve, the second is a linearized Euler equation, and the third is the monetary policy rule. For this section, we assume that the interest rate r_t is not observed. The unobserved exogenous shocks Δa_t and u_t are generated by the law

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

To solve the model analytically, in this section we make several simplifying assumptions. In particular, we assume that $\lambda = 0$, $\phi_x = 0$, $\phi_\pi = \frac{1}{b}$, and $\sigma^2 = 0$. The model then has six unknown scalar parameters: $\theta = (b, \kappa, \rho, \delta, \sigma_u^2, \sigma_a^2)$.

In the Supplement, we solve the model (1) under these restrictions to obtain

$$\begin{aligned} \begin{pmatrix} x_t \\ \pi_t \end{pmatrix} &= \begin{pmatrix} -\dfrac{b}{b+\kappa-\delta b} & \dfrac{b\rho}{b+\kappa-\rho b} \\ -\dfrac{b\kappa}{(b+\kappa-\delta b)(1-\delta b)} & \dfrac{b\kappa\rho}{(b+\kappa-\rho b)(1-b\rho)} \end{pmatrix} \begin{pmatrix} u_t \\ \Delta a_t \end{pmatrix} \\ &= C(\theta) \begin{pmatrix} u_t \\ \Delta a_t \end{pmatrix}. \end{aligned}$$

As we can see, the observed series x_t and π_t are weighted sums of two unobserved autoregressive processes with AR coefficients ρ and δ, where the weights depend on b and κ. It is relatively easy to see that if $0 < b < 1$, $\kappa > 0$, $\sigma_u^2 > 0$, $\sigma_a^2 > 0$, and $0 < \delta < \rho < 1$, then the six-dimensional parameter θ is point-identified.

Identification of the model fails when $\rho = \delta$. Indeed, there are two peculiarities in this case: first, if $\rho = \delta$, then u_t and Δa_t share the same autoregressive coefficient, and the dynamics of the observed series become insufficiently rich to disentangle the weight functions and separately identify b and κ. Second, the 2×2 matrix $C(\theta)$ becomes degenerate (of rank 1) at $\rho = \delta$. We show in the Supplement that at $\rho = \delta$, the parameter θ loses 2 degrees of identification. In this case, we can identify only a four-dimensional quantity: the two parameters ρ and δ, and the two functions $\frac{b}{b+\kappa-\rho b}\sqrt{\rho^2\sigma_a^2 + \sigma_u^2}$ and $\frac{\kappa}{1-\rho b}$, but not the parameters b, κ, σ_a^2, and σ_u^2 separately.

If $\rho = \delta$, underidentification precludes us from estimating the parameter θ consistently, and the usual asymptotic theory of maximum likelihood estimation does not apply. Even if $\rho \neq \delta$, when the difference $\rho - \delta$ is close to zero, we may have difficulty making reliable statistical inferences. In particular, the finite-sample size of many statistical tests may be quite far from the declared level and many conventional confidence sets may be misleading. To give a concrete example, consider the Wald statistic W for testing true hypothesis $H_0 : \theta = \theta_0$. According to the usual asymptotic theory of maximum likelihood, if $\rho \neq \delta$, then as the sample size T increases to infinity, the statistic W converges in distribution to a χ^2_6 under H_0. If, on the other hand, $\rho = \delta$, this convergence breaks down as the maximum likelihood estimator (MLE) is not consistent. Hence, the limit distribution of W experiences a discontinuity at $\rho = \delta$. Since the finite-sample distribution of W is continuous in the true parameter value, this implies that the convergence of W to a χ^2 distribution is not uniform in the parameter $\rho - \delta$ in a neighborhood of zero. Specifically, the closer $\rho - \delta$ is to zero, the larger a sample is required to achieve a given accuracy of approximation of the distribution of W by its asymptotic (χ^2_6) limit. This phenomenon is called weak identification.

To model the problems arising from weak identification, we can use a weak asymptotic embedding, considering a sequence of models such that $\rho = \delta + \frac{C}{\sqrt{T}}$, where C is a constant and T is the sample size. It is important to emphasize the conceptual essence of such an embedding: the researcher does not think that the parameters ρ and δ are changing with the sample size, but rather uses this embedding to obtain asymptotic approximations that reflect the trade-off between the proximity of the parameters ρ and δ and the quality of the classical asymptotic approximations. When examining asymptotic behavior along sequences of models with $\rho = \delta + \frac{C}{\sqrt{T}}$ as $T \to \infty$, we often find that some statistics, like W, have limiting distributions that differ from the χ^2 limits obtained under classical asymptotic theory. This reflects the sensitivity of those statistics to finiteness of information along some dimensions. If, however, we find a statistic that converges to the same χ^2 limit even under weak asymptotics, we call such a statistic *robust to weak identification*. Later, we will show that certain score statistics are robust in this sense.

Allowing the true parameter value to drift toward a point of nonidentification as the sample grows is one common way to model weak identification (see Andrews and Cheng (2012) on this), but there are other approaches. Under the approach of Stock and Wright (2000) for weakly identified generalized method of moments (GMM) models, for example, the objective function is modeled as indexed by the sample size and is taken to be asymptotically flat along some directions in the parameter space, thus not providing identification in the limit. This approach is not explicit about what parameter, if any, measures the proximity to identification failure; neither need it assume that there is any point of identification failure in finite samples. To cast the DSGE model discussed above into this framework, suppose for a moment that we know (or calibrate) the true values of $\rho \neq \delta$ so that ρ and δ are excluded from the parameter space. This does not solve the weak identification problem since the sample still contains limited information about the two weak directions if the calibrated values of ρ and δ are close. At the same time, the model is now point-identified over the whole parameter space.

The Supplement gives several stylized examples illustrating different types of weak identification that may arise in a DSGE context. In particular, we show how weak identification can arise from insufficiently rich dynamics of the observed process, for example, when autoregressive coefficients for several processes are close to each other or when moving average coefficients nearly cancel with autoregressive roots in an autoregressive moving average (ARMA) process. We also give a examples of a weakly identified vector autoregression (VAR) model and a nonlinear model with a weakly identified regime-switching mechanism.

3. Martingale methods in maximum likelihood

Let X_T be the data available at time T. To allow for the possibility of a weak identification embedding, 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 are generated by distribution $f_T(X_T;\theta_0)$, which may change as T grows. In general, we assume that $X_T = (x_{T,1},\dots,x_{T,T})$. Let $\mathcal{F}_{T,t}$ be a sigma algebra generated by the first t observations $X_{T,t} = (x_{T,1},\dots,x_{T,t})$. We assume that the log likelihood of the model,

$$\ell_T(X_T;\theta) = \log f_T(X_T;\theta) = \sum_{t=1}^{T} \log f_T(x_{T,t}|\mathcal{F}_{T,t-1};\theta),$$

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

Our main object of study will be the score function

$$S_T(\theta) = S_{T,T}(\theta) = \frac{\partial}{\partial\theta'}\ell_T(X_T,\theta) = \sum_{t=1}^{T}\frac{\partial}{\partial\theta'}\log f_T(x_{T,t}|\mathcal{F}_{T,t-1};\theta),$$

and we take $s_{T,t}(\theta) = S_{T,t}(\theta) - S_{T,t-1}(\theta) = \frac{\partial}{\partial\theta'}\log f_T(x_{T,t}|\mathcal{F}_{T,t-1};\theta)$ to denote the increment of the score. Under the assumption that we have correctly specified the model, we have that $E(s_{T,t}(\theta_0)|\mathcal{F}_{T,t-1}) = 0$ almost surely. This in turn implies that for each T, the score taken at the true parameter value, $S_{T,t}(\theta_0)$, is a martingale with respect to filtration $\mathcal{F}_{T,t}$. This is a generalization of the first informational equality due to Silvey (1961).

Similarly, the second informational equality also generalizes to the dependent case. This equality states that we can calculate the (theoretical) Fisher information, $\mathcal{I}_T(\theta_0)$, either as the expectation of the negative Hessian of the log likelihood or as the expectation of the outer product of the score. Fisher 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 the second informational equality suggests two different ways of estimating it that are asymptotically equivalent in the classical context. To generalize the second informational equality to the dynamic context, following Barndorff-Nielsen and

Sorensen (1991), we introduce two measures of information based on observed quantities. The first is the *observed information* and is equal to the negative Hessian of the log likelihood,

$$I_T(\theta) = -\frac{\partial^2}{\partial\theta\,\partial\theta'}\ell_T(X_T;\theta) = \sum_{t=1}^{T} i_{T,t}(\theta),$$

where $i_{T,t}(\theta) = -\frac{\partial^2}{\partial\theta\,\partial\theta'}\log f_T(x_{T,t}|\mathcal{F}_{T,t-1};\theta)$. The second is the *incremental observed information* and is equal to the quadratic variation of the score,

$$J_T(\theta) = \big[S_T(\theta)\big] = \sum_{t=1}^{T} s_{T,t}(\theta)s'_{T,t}(\theta),$$

where as before $s_{T,t}(\theta)$ is the increment of $S_T(\theta)$. Both observed measures $I_T(\theta)$ and $J_T(\theta)$ are unbiased estimates of the (theoretical) Fisher information for the whole sample: $\mathcal{I}_T(\theta_0) = E(I_T(\theta_0)) = E(J_T(\theta_0))$. 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 $A_{T,t}(\theta_0)$ is a martingale with respect to $\mathcal{F}_{T,t}$. Specifically, the increment of $A_{T,t}(\theta_0)$ is

$$a_{T,t}(\theta_0) = A_{T,t}(\theta_0) - A_{T,t-1}(\theta_0) = s_{T,t}(\theta_0)s'_{T,t}(\theta_0) - i_{T,t}(\theta_0),$$

and a simple argument gives us that $E(a_{T,t}|\mathcal{F}_{T,t-1}) = 0$ almost surely (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 independent and identically distributed (i.i.d.) case, for example, the law of large lumbers implies that $\frac{1}{T}I_T(\theta_0) \xrightarrow{p} -E(\frac{\partial^2}{\partial\theta\,\partial\theta'}\log f(x_t,\theta_0)) = \mathcal{I}_1(\theta_0)$ and $\frac{1}{T}J_T(\theta_0) \xrightarrow{p} E(\frac{\partial}{\partial\theta'}\log f(x_t,\theta_0)\frac{\partial}{\partial\theta}\log f(x_t,\theta_0)) = \mathcal{I}_1(\theta_0)$. 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 the 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 MLE is asymptotically normal and $J_T(\theta_0)^{1/2}(\hat{\theta} - \theta_0) \Rightarrow N(0,1)$.

We depart from this classical approach in that we consider weak identification. We find that in weakly identified models, 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 $J_T(\theta_0)$.

Example 1. To illustrate this nonequivalence in a simple example, suppose we observe data Y_t for $t \in \{1,\dots,T\}$, generated by the model

$$Y_t = (\pi+\beta)Y_{t-1} + e_t - \pi e_{t-1}, \quad e_t \sim \text{i.i.d. } N(0,1). \tag{3}$$

The true value of the parameter $\theta_0 = (\beta_0, \pi_0)'$ satisfies the 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 the initial condition will not matter asymptotically. One can rewrite the model as $(1 - (\pi + \beta)L)Y_t = (1 - \pi L)e_t$. It is easy to see that if $\beta_0 = 0$, then the parameter π is not identified. Andrews and Cheng (2012) modeled weak identification using the drifting parameter value $\beta_0 = \frac{C}{\sqrt{T}}$, leading to the parameter π being weakly identified.

Consider the normalization matrix $K_T = \text{diag}(1/\sqrt{T}, 1)$. Then

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

where Σ is a positive-definite matrix, while ξ and η are two Gaussian random variables.[2] As we can see, the difference between the two information matrices is asymptotically nonnegligible compared with the information measure $J_T(\theta_0)$.

The Supplement contains several examples of weakly identified models. For all of them, we observe the same phenomenon: the appropriately normalized quadratic variation of the score J_T converges in probability to a positive-definite matrix, while the Hessian normalized in the same way converges weakly to a random matrix. White (1982) shows that the two measures of information may differ if the likelihood is misspecified. As our examples show, even if the model is correctly specified these two measures may differ substantially if identification is weak. This result is quite different from that of White (1982). In particular, correct specification implies that $EA_T(\theta_0) = 0$, and it is this restriction that is tested by White's information matrix test. In contrast, weak identification in correctly specified models is related to $A_T(\theta_0)$ being substantially volatile relative to $J_T(\theta_0)$ while maintaining the assumption that $EA_T(\theta_0) = 0$. Correct specification can still be tested by comparing the realized value of $A_T(\theta_0)$ to the metric implied by a consistent estimator of its variance. One may potentially create a test for weak identification based on a comparison of A_T with J_T, though this is beyond the scope of the present paper. We will, however, treat nonpositive-definiteness of the Hessian as an informal sign of weak identification.

4. Test for full parameter vector

In this section, we suggest tests for a simple hypothesis on the full parameter vector, $H_0 : \theta = \theta_0$, which are robust to weak identification. We introduce our first assumption.

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

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

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

[2] Details can be found in the Supplement.

Discussion of Assumption 1. Assumption 1(a) is a classical infinitesimality (or limit negligibility) condition. It requires that no single observation matter too much asymptotically, and holds quite generally in stationary models. Assumption 1(b) imposes the ergodicity of the quadratic variation $J_T(\theta_0)$ of martingale $S_T(\theta_0) = S_{T,T}(\theta_0)$, which rules out some potentially interesting models including persistent (unit root) processes and nonergodic models. A key aspect of Assumption 1 is that we impose no restriction on the form of the sequence of normalizing matrices K_T. In particular, while in strongly identified models we can generally take $K_T = \frac{1}{\sqrt{T}}\mathrm{Id}_k$, in weakly identified models we will typically need to take some directions of K_T to be constant, or even growing with T, to obtain an appropriate normalization.

Assumption 1 holds for the analytically solved DSGE model discussed in Section 2 and for Example 1 above. It also holds in all the weakly identified models we examine in the Supplement. Under this assumption, we obtain the following theorem as a direct corollary of the multivariate martingale central limit theorem (see Theorem 8, Chapter 5 in Liptser and Shiryayev (1989)).

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

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

$$\mathrm{LM}_e(\theta_0) = S_T(\theta_0) \mathcal{I}_T(\theta_0)^{-1} S_T(\theta_0) \Rightarrow \chi^2_k, \tag{5}$$

where $k = \dim(\theta_0)$.

We consider two formulations of the well known LM statistic in equations (4) and (5), one using observed incremental information $J_T(\theta_0)$ and the other using the (expected) Fisher information $\mathcal{I}_T(\theta_0)$. Theorem 1 shows that pairing either of these statistics with χ^2_k critical values results in a weak identification-robust test. The two statistics are asymptotically equivalent under the null provided Assumption 1 holds, but may have different finite-sample performance, and we find in simulations (see Section 7) that $\mathrm{LM}_e(\theta_0)$ controls size somewhat better. On the other hand, the statistic $\mathrm{LM}_o(\theta_0)$ has two advantages. First, in many cases, calculating $J_T(\theta_0)$ is much more straightforward than calculating $\mathcal{I}_T(\theta_0)$, particularly when we do not have an analytic expression for the likelihood. Second, if we weaken Assumption 1(b) to require only that Σ be an almost surely positive-definite *random* matrix, then (4) still holds while (5) does not. Hence (4), unlike (5), has the additional advantage of being robust to nonergodicity. Statistical examples of nonergodic models can be found in Basawa and Koul (1979).

Unlike the classical maximum likelihood (ML) Wald and likelihood ratio (LR) tests, the derivation of the asymptotic distribution of the LM statistics (4) and (5) uses no assumptions about the strength of identification. It is important to note, however, that the LM statistic calculated with other estimators of the Fisher information (for example, $I_T(\theta_0)$) is not necessarily robust to weak identification. It is also unwise to estimate the information matrix using an estimator of θ, that is, to use $J_T(\hat{\theta})$. All of these alternative formulations deliver asymptotically equivalent tests in strongly identified models, but this equivalence fails under weak identification.

A REMARK ON POINT VERSUS WEAK IDENTIFICATION. Assumption 1(b) rules out locally nonidentified models by assuming that Σ is positive definite. In ML models, it is usually possible to check local point identification by checking the nondegeneracy of the Fisher information. The corresponding literature for DSGE models includes Komunjer and Ng (2011) and Iskrev (2010). If one wants to test the full parameter vector at a point of nonidentification, under the null there exists a nondegenerate linear transformation of the score such that a subvector of the transformed score is identically zero while the rest has nondegenerate quadratic variation. If Assumption 1 holds for the nonzero part of the transformed score, our LM tests (replacing the inverse with the Moore–Penrose pseudo-inverse) are asymptotically χ^2-distributed with reduced degrees of freedom.[3] See Andrews (1987) for a discussion of related issues.

5. TEST FOR A SUBSET OF PARAMETERS

In applied economics, it is very common to report separate confidence intervals for each one-dimensional subparameter in the 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 one-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 subparameters, 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 the proposed weak identification-robust LM tests of the full parameter vector, we have the option to produce robust confidence sets for subparameters via the projection method. This approach has been used many times in the literature, for example, by Dufour and Taamouti (2005) for weak instrumental variables (IV) and by DKK for DSGE. The typical DSGE model has a large number of parameters to estimate (often between 20 and 60), however, which makes the projection method less attractive as the degree of conservativeness may be very high, rendering the resulting confidence sets less informative. Below, we introduce an alternative procedure that has better power properties than the projection method but can only be applied under additional assumptions.

5.1 *LM statistic for composite hypotheses*

Assume that $\theta = (\alpha', \beta')'$ and we are interested in constructing a robust test of the hypothesis $H_0 : \beta = \beta_0$, while treating α as a nuisance parameter. We consider the same LM statistics as defined in (4) and (5) and evaluated at $\theta = (\hat{\alpha}, \beta_0)$, where $\hat{\alpha}$ is the restricted MLE, that is, $\hat{\alpha} = \arg\max_\alpha \ell(\alpha, \beta_0)$. Denoting our subset tests by $\widetilde{\mathrm{LM}}_o(\beta_0)$ and $\widetilde{\mathrm{LM}}_e(\beta_0)$, we have that

$$\widetilde{\mathrm{LM}}_o(\beta_0) = \mathrm{LM}_o(\hat{\alpha}, \beta_0) = S'_\beta \big(J_{\beta\beta} - J_{\beta\alpha} J_{\alpha\alpha}^{-1} J'_{\beta\alpha}\big)^{-1} S_\beta\big|_{\theta=(\hat{\alpha},\beta_0)}, \tag{6}$$

[3] We are grateful to an anonymous referee for pointing this out.

where $S_T(\theta) = (S_\alpha(\theta)', S_\beta(\theta)')'$ and $J(\theta) = \begin{pmatrix} J_{\alpha\alpha} & J_{\alpha\beta} \\ J'_{\alpha\beta} & J_{\beta\beta} \end{pmatrix}$ are the natural partitions of the score and observed information. Statistic $\widetilde{\mathrm{LM}}_e(\beta_0)$ can be defined analogously using statistic $\mathrm{LM}_e(\theta_0)$.

The classical theory of maximum likelihood considers two LM tests for such a setting: Rao's score test and Neyman's $C(\alpha)$ test. Rao's score test is based on the statistic $\mathrm{Rao} = \frac{1}{T} S_T(\hat{\theta}_0)' \mathcal{I}(\hat{\theta}_0)^{-1} S_T(\hat{\theta}_0)$, where $\hat{\theta}_0$ is the restricted ML estimator, while Neyman's $C(\alpha)$ test was developed as a locally asymptotically most powerful (LAMP) test for composite hypotheses in the classical ML framework. If the classical ML assumptions are satisfied, both statistics have an asymptotic $\chi^2_{k_\beta}$ distribution, and, in fact, Kocherlakota and Kocherlakota (1991) show that the two statistics are asymptotically equivalent. One can also see that our proposed statistics are asymptotically equivalent to both Rao's score and Neyman's $C(\alpha)$ if the classical ML assumptions are satisfied, and hence that our test does not lose power compared to the classical tests if the model is strongly identified.

The approach we take in this paper differs from that of Stock and Wright (2000). In particular, rather than minimizing the LM statistic over the nuisance parameter α as in Stock and Wright (2000) and the projection method, we instead plug in the restricted ML estimate. One may show in a linear weak IV model that plugging in the restricted MLE for strongly identified nuisance parameters leads to a χ^2 limiting distribution, while minimizing the LM statistic does not.

5.2 *Robust tests with strong nuisance parameters*

The critical issue in the literature on robust testing is whether α is weakly or strongly identified. In this section, we provide conditions that guarantee that the subset LM tests will be asymptotically valid in models with strongly identified nuisance parameters. We begin by adapting Bhat's (1974) result to establish the consistency and asymptotic normality of the MLE. Let $A_{\alpha\alpha,T} = J_{\alpha\alpha,T} - I_{\alpha\alpha,T}$, where the last two quantities are the submatrices 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*[4] *with $K_{\alpha,T}$ and $K_{\beta,T}$ the submatrices of K_T corresponding to α and β, respectively. Furthermore,*

(a) $K_{\alpha,T} A_{\alpha\alpha,T} K_{\alpha,T} \xrightarrow{p} 0$,

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

$$\sup_{\|K_{\alpha,T}^{-1}(\alpha_1 - \alpha_0)\| < \delta} \left\| K_{\alpha,T} \big(I_{\alpha\alpha}(\alpha_1, \beta_0) - I_{\alpha\alpha}(\alpha_0, \beta_0) \big) K_{\alpha,T} \right\| \xrightarrow{p} 0,$$

(c) $\hat{\alpha}(\beta_0)$ *is such that* $K_{\alpha,T}^{-1}(\hat{\alpha} - \alpha_0) = O_p(1)$.

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

$$K_{\alpha,T}^{-1}(\hat{\alpha} - \alpha_0) = K_{\alpha,T}^{-1} J_{\alpha\alpha,T}^{-1} S_{\alpha,T} + o_p(1) \Rightarrow N\big(0, \Sigma_{\alpha\alpha}^{-1}\big). \tag{7}$$

[4]Lemma 1 continues to hold if we replace the diagonality assumption on K_T by the requirement that K_T be block-diagonal with blocks $K_{\alpha,T}$ and $K_{\beta,T}$.

Discussion of Assumption 2. Assumption 2(a) implies that $K_{\alpha,T} I_{\alpha\alpha,T} K_{\alpha,T} \xrightarrow{p} \Sigma_{\alpha\alpha}$ and, hence, that the two observed information matrices for α are the same asymptotically. We mentioned a condition of this nature in our discussion of weak identification in Section 3. One approach to checking Assumption 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

$$A_{\alpha\alpha,T} = \sum_{t=1}^{T} \frac{1}{f_T(x_{T,t}|\mathcal{F}_{T,t-1}, \theta_0)} \frac{\partial^2}{\partial\alpha\,\partial\alpha'} f_T(x_{T,t}|\mathcal{F}_{T,t-1}, \theta_0).$$

If the terms $\frac{1}{f_T(x_{T,t}|\mathcal{F}_{T,t-1},\theta_0)} \frac{\partial^2}{\partial\alpha\,\partial\alpha'} f_T(x_{T,t}|\mathcal{F}_{T,t-1}, \theta_0)$ 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(c) is a high-level assumption on the behavior of the restricted MLE. If $K_{T,\alpha}$ is decreasing to zero, then this assumption requires that the restricted MLE for the nuisance parameter α be consistent at a particular rate under the null. Such consistency can be obtained using standard arguments for strongly identified models, for example, by appealing to uniform convergence of the objective function together with identification of α. Assumption 2(b) is an assumption on the smoothness of the log likelihood.

Assumption 3. *Consider the sequence of martingales*

$$M_T = \big(S_T(\theta_0)', \operatorname{vec}\big(A_{\alpha\beta,T}(\theta_0)\big)'\big)' = \sum_{t=1}^{T} m_{t,T}.$$

Assume that there exists a sequence of nonstochastic diagonal matrices $K_{M,T}$ such that

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

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

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

$$\Lambda_{\alpha_i\alpha_j\beta_n} = \sum_{t=1}^{T} \frac{1}{f_T(x_{T,t}|\mathcal{F}_{T,t-1}, \theta_0)} \cdot \frac{\partial^3 f_T(x_{T,t}|\mathcal{F}_{T,t-1}, \theta_0)}{\partial\alpha_i\,\partial\alpha_j\,\partial\beta_n}.$$

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 assumption.

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

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

Discussion of Assumption 4. Assumption 4(b) and (c) state that higher order likelihood 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 a law of large numbers.

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

Example 1 (Continued). In the Supplement, we show that Assumptions 2, 3, and 4 hold in the ARMA(1, 1) model with nearly canceling roots when 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.

6. Suggestions for applied researchers

Below we highlight some practical details concerning testing and confidence set construction that are particularly relevant for applied researchers interested in using the tests discussed in this paper. First, one tractable approach to calculating the score in models where the likelihood is not available analytically is to approximate derivatives by considering appropriately scaled small differences (i.e., numerical derivatives). While the correct step size for such differences is typically not obvious, in the DSGE application studied in this paper, we have found that our results are generally insensitive to the choice of step size (though this will, of course, not be the case universally). The results discussed in the simulation section below, for example, were generated using finite differences with steps of size 10^{-6}, but considering steps of size 10^{-5} instead yields the same results.

Second, calculating the observed incremental information $J_T(\theta_0)$ is typically quite straightforward in linear, Gaussian models. Unfortunately, calculating the theoretical Fisher information $\mathcal{I}_T(\theta_0)$ can be considerably more involved, especially in models where the likelihood is not available analytically. One way to approximate the theoretical Fisher information is by averaging draws of the observed information I_T or observed incremental information J_T over a large number of simulations, but calculating the Fisher information in this way can be slow. In our simulations, we instead use an approach suggested by Iskrev (2008). Specifically, we first calculate the information matrix with respect to the parameters of the DSGE model's state-space representation and then use this information matrix, together with the derivatives of the state-space parameters with respect to the structural parameters (which we evaluate analytically, though approximating them numerically gives the same results), to obtain the information matrix for the nine model parameters. For further details and additional references, see Iskrev (2008) and Iskrev (2010). There are packages available for Matlab that can be used to evaluate the theoretical information matrix for the state-space parameters in linear models with Gaussian shocks. In particular, we use the e4 time-series toolbox for Matlab (see Jerez, Casals, and Sotoca (2011)).

The third suggestion is related to construction of confidence sets by inverting the tests proposed in this paper. So as to calculate a 95% LM_o confidence set for the parameter β, for example, we need to collect all values β_0 such that $H_0 : \beta = \beta_0$ is not rejected by an LM_o test with size 5%. How best to do this in practice depends on the context. For cases—like many DSGE applications—where the researcher specifies a bounded parameter space, the simplest approach may be to take draws at random from the parameter space for β, storing those values that are not rejected. We implement this approach to calculate LM_o-based confidence sets in the simulation section below. One may alternatively evaluate the test on a grid of points and record those values that are not rejected, though this may be very computationally costly when β is high dimensional. To create a projection-method confidence interval for a component β_i of β, we can take the upper and lower bounds to be the largest and smallest values β_i consistent with nonrejected values of β, which corresponds to projecting the convex hull of the nonrejected values of β on subspace corresponding to β_i.

Finally, our results allow a researcher to plug in the restricted MLE for well identified nuisance parameters, but to apply this approach one needs to know that particular parameters are strongly identified. This is a considerable problem in many DSGE models, and we are unaware of any test applicable to DSGE models that can discriminate between strongly and weakly identified parameters. In particular we are unaware of a pretest that, if we plug in the restricted MLE for those nuisance parameters that the pretest indicates are strongly identified, ensures that the resulting test controls the size of the two-step procedure. Absent formal results, we are left to rely on more indirect evidence on which parameters may be well identified. One indirect approach based on our results is to check whether a submatrix of the Hessian $I_T(\theta_0)$ corresponding to potentially strongly identified nuisance parameters is positive-definite with high probability. There is a common perception that in many models, parameters related to the variance and persistence of exogenous shocks, as well as steady-state parameters, may be relatively well identified provided the other model parameters are known.[5] Simulation results in the next section seem to bear this out in a small-scale DSGE model. When one is uncertain about the strength of identification for a given parameter, one can always err on the side of caution and project over that parameter, but minimizing the number of nuisance parameters to be projected over yields more powerful tests.

7. A small-scale DSGE model: Simulation results

We have a number of simulation results that both support our theoretical results and suggest directions for further research. We consider a simple DSGE model based on Clarida, Gali, and Gertler (1999). We assume that the econometrician observes a sample $\{(\pi_t, x_t, r_t), t = 1, \ldots, T\}$ from a data-generating process satisfying the (log-linearized) equilibrium conditions (1) and (2). The model has 10 parameters: the discount rate b, Calvo parameter κ, the Taylor rule parameters ϕ_x, ϕ_π, and λ, and the parameters describing the evolution of the exogenous variables. We will treat parameter $b = 0.99$ as known (and calibrated to its true value). We assume that the

[5]We thank Frank Schorfheide for bringing this to our attention.

TABLE 1. True parameter values for simulations.

	ϕ_x	ϕ_π	λ	ρ	δ	κ	σ_a	σ_u	σ
Calibrated value	2.28	2.02	0.898	0.85	0.103	0.1	0.325	0.265	0.556
Parameter space lower bound	0	0	0	−0.99	−0.99	0	0	0	0
Parameter space upper bound	10	10	0.99	0.99	0.99	1	1	1	1

econometrician is concerned with inference on the remaining nine parameters $\theta = (\phi_x, \phi_\pi, \lambda, \rho, \delta, \kappa, \sigma_a, \sigma_u, \sigma)'$. Note that unlike in Section 2, here we take the interest rate r_t to be observable and do not restrict the parameters other than b.

For our simulation exercise, we draw samples from the model with parameters calibrated to ML estimates obtained using demeaned U.S. macro data from Smets and Wouters (2007). The ML estimate of the parameter ρ is very close to 1, so since robustness to unit roots lies beyond the scope of the present paper, for our simulations we will instead use the smaller value $\rho = 0.85$. Likewise, the ML estimate for κ lies quite close to 0, which is the boundary of the parameter space for this parameter. To ensure that parameter-on-the-boundary issues do not greatly affect the distribution of classical test statistics, we increase the value of this parameter, taking $\kappa = 0.1$. The baseline values of parameters used in the simulations are reported in Table 1. The structural parameters are point-identified at this parameter value. We generate samples of size 300 from this model and then discard the first 100 observations, using only the last 200 for the remainder of the analysis.

7.1 *Properties of classical ML testing*

We begin by examining the behavior of the classical maximum-likelihood-based statistics. Histograms for the ML estimator[6] show that the marginal distributions of the estimates for several parameters depart substantially from a normal distribution. We consider four variations on the Wald statistic for testing the simple hypothesis $H_0 : \theta = \theta_0$, where θ_0 is the true value, corresponding to different estimators of the asymptotic variance, $\hat{V}$, used in the quadratic form $(\hat{\theta} - \theta_0)'\hat{V}^{-1}(\hat{\theta} - \theta_0)$. In particular, Wald $(I_T(\hat{\theta}))$ uses the inverse of the observed information, evaluated at $\hat{\theta}$, to estimate the asymptotic variance. Wald $(I_T(\theta_0))$, on the other hand, evaluates the observed information at the true parameter value. Likewise, Wald $(J_T(\hat{\theta}))$ and Wald $(J_T(\theta_0))$ use J_T^{-1} as the estimator of the asymptotic variance, calculated at $\hat{\theta}$ and θ_0, respectively. Under the usual strong identification assumptions for ML, all of these statistics should have a χ^2_9 distribution asymptotically. In simulation, however, the distribution of these statistics appears quite far from a χ^2_9. Table 2 lists sizes for nominal 5% and 10% tests (based on 2500 simulations), and shows that all versions of the Wald test we consider severely overreject. Taken together, these results strongly suggest that the usual approaches to ML estimation and inference are poorly behaved when applied to this DSGE model.

[6]Available from the authors by request.

TABLE 2. Simulated size of Wald tests for the nine-dimensional hypothesis $H_0 : \theta = \theta_0$; based on 2500 simulations.

	Wald ($I_T(\theta_0)$)	Wald ($I_T(\hat{\theta})$)	Wald ($J_T(\theta_0)$)	Wald ($J_T(\hat{\theta})$)
Size of 5% test	39.16%	42.36%	40.24%	40.44%
Size of 10% test	43.2%	47.44%	45.72%	45.88%

7.2 *Behavior of the information matrix*

In Section 3, we associated weak identification with the difference between two information measures $A_T(\theta_0)$ being large compared to $J_T(\theta_0)$. Note 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 $I_T(\theta_0) = J_T(\theta_0) - A_T(\theta_0)$ will be positive-definite for almost all realizations of the data. We can check positive-definiteness of $I_T(\theta_0)$ directly in simulations. Considering the observed information evaluated at the true value, we find that it has at least one negative eigenvalue in over 47% of simulation draws (based on 2500 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 poor behavior of ML estimation in this model. In line with the conjecture discussed above that the persistence and variance parameters may be well identified if we know the structural parameters $(\phi_x, \phi_\pi, \lambda, \kappa)$, we find that the observed information for the five parameters $(\rho, \delta, \sigma_a, \sigma_u, \sigma)$ alone is positive-definite in all simulation draws.

7.3 *Size of the LM tests*

We now turn to the weak identification-robust statistics discussed earlier in this paper. Under appropriate assumptions, we have that $\mathrm{LM}_o(\theta_0) \Rightarrow \chi^2_9$ and $\mathrm{LM}_e(\theta_0) \Rightarrow \chi^2_9$, where $\mathrm{LM}_o(\theta_0)$ is the LM statistic using the observed incremental information $J_T(\theta_0)$ and $\mathrm{LM}_e(\theta_0)$ is calculated with the theoretical Fisher information $\mathcal{I}_T(\theta_0)$. In Figure 1, we plot the cumulative distribution functions (CDFs) of the simulated distributions of $\mathrm{LM}_o(\theta_0)$ and $\mathrm{LM}_e(\theta_0)$ together with a χ^2_9. Table 3 reports the size of the LM tests. Two points are clear from these results: first, though our tests based on the LM statistics are not exact, the χ^2 approximation is very good for LM_e and reasonable for LM_o. Second, the LM_e statistic has somewhat better finite-sample properties.

We next consider the size of the two LM statistics for testing subsets of parameters. Specifically, as before, we consider a partition of the parameter vector, $\theta = (\alpha', \beta')'$, and consider the problem of testing $H_0 : \beta = \beta_0$, treating α as a nuisance parameter.

As discussed in Section 5, an important issue is whether the nuisance parameter α is weakly or strongly identified. While we are unaware of any formal tests for identification strength in DSGE models that ensure size control when used as pretests, there is a common perception that, fixing structural parameters like ϕ_x, ϕ_π, λ, and κ, the parameters controlling the persistence and the variance of shocks will be well identified. Since this is consistent with our results from comparing different information measures, we treat these parameters as strongly identified.

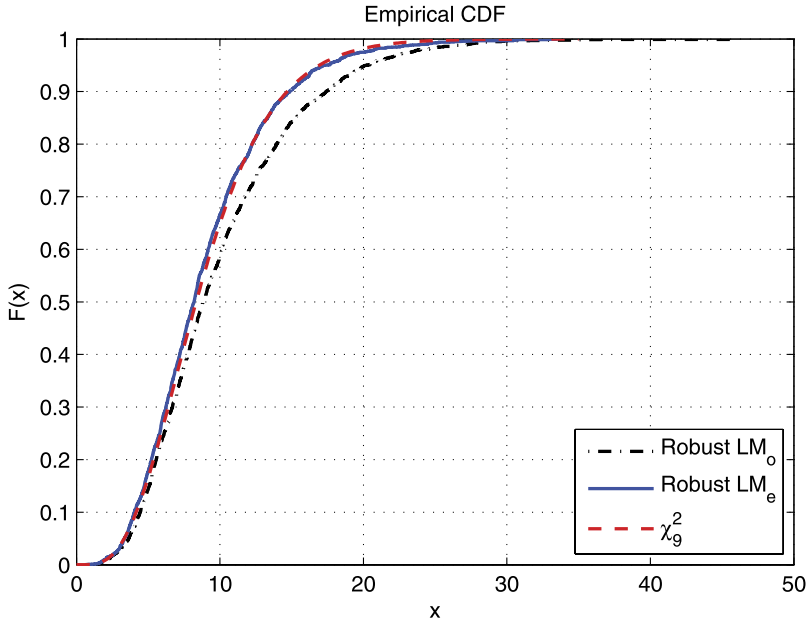

FIGURE 1. CDF of simulated LM statistics introduced in Theorem 1 compared to χ^2_9.

TABLE 3. Simulated size (based on 1000 simulations) of a test for the full parameter vector and for five tests of composite hypotheses $H_0 : \beta = \beta_0$, treating all other parameters as nuisance parameters.

		LM$_o$		LM$_e$	
	Tested Parameters	5%	10%	5%	10%
	All parameters	9%	15.3%	4.5%	9.1%
(∗1)	$\beta = (\phi_x, \phi_\pi, \lambda, \kappa)$	5.9%	11.1%	4.4%	8.1%
(∗2)	$\beta = (\phi_x, \phi_\pi, \lambda, \kappa, \rho)$	6.3%	11.5%	5.1%	9.1%
(∗3)	$\beta = (\phi_x, \phi_\pi, \lambda, \kappa, \delta)$	5.9%	11.6%	4.1%	8.6%
(∗4)	$\beta = (\phi_x, \phi_\pi, \lambda, \kappa, \sigma_a)$	5.9%	10.8%	4.4%	8.2%
(∗5)	$\beta = (\phi_x, \phi_\pi, \lambda, \kappa, \sigma_u)$	5.9%	11.2%	4.0%	8.1%
(∗6)	$\beta = (\phi_x, \phi_\pi, \lambda, \kappa, \sigma)$	7.2%	13%	4.9%	9.6%

Note: Statistic LM$_o$ refers to the LM test using observed incremental information and statistic LM$_e$ uses theoretical Fisher information, and in both cases we plug in the restricted MLE for nuisance parameters.

We consider testing six different composite hypotheses (corresponding to cases (∗1)–(∗6) in Table 3): a hypothesis on the four structural parameters $(\phi_x, \phi_\pi, \lambda, \kappa)$ and five hypotheses on these four parameters plus each of the other five parameters taken one at a time, $(\phi_x, \phi_\pi, \lambda, \kappa, \rho)$, $(\phi_x, \phi_\pi, \lambda, \kappa, \delta)$, and so forth. In each case we follow the approach discussed in Section 5 and plug in the restricted MLE for the parameters not under test, reducing the critical value appropriately. Our simulation results, reported in Table 3, are consistent with the assumption that the parameters $(\rho, \delta, \sigma_a, \sigma_u, \sigma)$ are strongly identified. In particular, we see that all the tests we consider for composite hy-

TABLE 4. 95% LM_o confidence intervals for parameters based on single draw of simulated data, where we treat the parameters $(\rho, \delta, \sigma_a, \sigma_u, \sigma)$ as well identified and project over the other parameters.

Level	ϕ_x	ϕ_π	λ	ρ	δ	κ	σ_a	σ_u	σ
Lower	1.04	0.58	0.76	0.74	0.04	0	0.28	0.24	0.46
Upper	9.97	8.88	0.97	0.91	0.46	0.18	0.50	0.34	0.56

potheses control size fairly well, though the size control of the LM_e tests is again somewhat better.

7.4 *Calculation of confidence sets*

Despite weak identification, we can produce informative confidence sets. To illustrate this point, we take one random draw from the model, treat it as a sample, and report LM_o confidence intervals for each of our nine parameters separately in Table 4.

To calculate these one-dimensional confidence intervals, we follow the approach discussed in Section 6, and first form four- and five-dimensional confidence sets by inverting the LM_o tests for the six composite hypotheses corresponding to $(*1)$–$(*6)$ in Table 3, that is, for each group of parameters, we collect all values of β_0 such that the corresponding hypotheses $H_0 : \beta = \beta_0$ are not rejected. For example, in case $(*1)$, we construct a joint four-dimensional confidence set for parameters $(\phi_x, \phi_\pi, \lambda, \kappa)$. For each group of tested parameters, we take $5 \cdot 10^4$ draws uniformly at random over the parameter space for β formed by the Cartesian product of the one-dimensional parameter spaces given in Table 1, and keep those draws that are not rejected by the LM_o test that plugs in the restricted MLE for the nuisance parameters (all parameters other than β). By projecting the (four-dimensional) convex set obtained for the case $(*1)$ on the subspace corresponding to each parameter separately, we obtain one-dimensional confidence sets for each of the parameters ϕ_x, ϕ_π, λ, and κ. To obtain one-dimensional confidence sets for the remaining five parameters ρ, δ, σ_a, σ_u, and σ, we project the corresponding five-dimensional confidence sets obtained for cases $(*2)$–$(*6)$ on the subspace corresponding to the parameter of interest. We can see that while the confidence intervals for many parameters are wide, in all instances they exclude some values and in most cases they cover only a small portion of the parameter space.

7.5 *Alternative weak identification-robust methods*

Issues of weak identification in DSGE models have recently attracted the attention of econometricians, and several weak identification-robust methods for DSGE models have been suggested independently by Dufour, Khalaf, and Kichian (2013) (DKK), Guerron-Quintana, Inoue, and Kilian (2013) (GQIK), and Qu (forthcoming). It is important to note that DKK and Qu focus primarily on testing the full parameter vector, while GQIK allow one to concentrate out strongly identified nuisance parameters. None of the competing papers offers procedures to determine which specific parameters are

strongly identified. They all use projection for testing with weak nuisance parameters or parameters whose identification strength is unknown.

Our method differs from the three approaches mentioned above in that it is valid in a general ML framework with potentially weak identification and is not restricted to log-linearized DSGE models. The LM statistics we propose can be used whenever we can evaluate the likelihood function. In contrast, the three approaches above are specially designed for log-linearized DSGE models that can be written as linear expectation equations. In general, these methods cannot be applied to the nonlinear DSGE models that are increasingly popular; see, for example, Fernández-Villaverde and Rubio-Ramírez (2011). Though the range of nonlinear DSGE models for which one can differentiate the likelihood function is quite limited at present, the number of such models is growing; see, for example, Amisano and Tristani (2011).

The method closest to ours is the LM test suggested by Qu (forthcoming) for log-linearized DSGE models with normal errors. Qu (forthcoming) notices that in large samples, the Fourier transforms of the observed data at different frequencies are approximately independent Gaussian random variables with variance equal to the spectrum of the observed series; this allows him to write an approximate likelihood for the data in a very elegant way and to discuss the properties of the likelihood analytically. His statistic is almost the same as our statistic $\mathrm{LM}_e(\theta_0)$ for testing the full parameter vector, the main difference being that Qu (forthcoming) uses an approximate likelihood, while we use the exact likelihood. Hence, we expect that the two statistics applied to a log-linearized DSGE model with normal errors should be very close provided Qu's approximate likelihood is well behaved.

GQIK consider models with a linear state-space representation and assume that the coefficients of the state-space representation, $Y = Y(\theta)$, are either strongly identified or not identified at all, while no assumption is made on the identification of the structural parameters θ. For testing a hypothesis $H_0 : \theta = \theta_0$ about the structural parameter vector, GQIK suggest testing the hypothesis $\widetilde{H}_0 : Y = Y(\theta_0)$ about the reduced-form parameter, using the classical LR statistic and the usual χ^2 critical values with degrees of freedom equal to the dimensionality of the identified reduced-form parameter. The assumption of GQIK that the reduced-form parameters are strongly identified seems quite problematic in some DSGE applications and no test is available to check it. Schorfheide (2010) provides an example in which weak identification of the structural parameters leads to weakly identified reduced-form parameters. Unlike the tests suggested in this paper, the LR test proposed by GQIK is typically asymptotically inefficient under strong identification, since the dimension of the reduced-form parameter is usually higher than that of the structural parameter. GQIK also suggest a test based on Bayes factors, which we do not discuss here as it is less directly comparable to our approach.

DKK propose a limited information approach based on a set of exclusion restrictions implied by a system of linear expectation equations, which they then test using a seemingly unrelated regression-based (SUR-based) F-statistic in the spirit of Stock and Wright (2000). Advantages of this approach are that a researcher has the freedom to choose which restrictions he or she wishes to use for inference, and that it does not require distributional assumptions on the error term and hence is robust to misspecifica-

tion. A disadvantage of the method is its limited ability to accommodate latent state variables. Furthermore, this limited information test may be expected to have lower power than full-information methods if the model is correctly specified. DKK also suggested a full-information ML method based on a VAR approximation to the DSGE solution, but the authors seem to prefer and advocate their limited information approach, so we focus on this method.

7.6 *Power comparisons with alternative methods*

Here we compare the power of the alternative approaches to that of the proposed LM tests. As the alternative approaches deal primarily with testing the full parameter vector, we will focus on this case.

Table 5 reports actual size, while Figure 2 shows (non-size-corrected) power curves for 5% tests based on the statistics $\mathrm{LM}_o(\theta_0)$ and $\mathrm{LM}_e(\theta_0)$, a version of Qu's (forthcoming) LM test, the LR test introduced in GQIK, and the limited information (LI) test of DKK. Implementation details are discussed below. Power is calculated for alternatives that entail a change in one element of the parameter vector while the other elements remain at their null values. The label on each subplot denotes the parameter whose value changes under the alternative.

First, we consider Qu's (forthcoming) frequency-domain LM test.[7] Initial simulations showed that this test tended to overreject at some parameter values and that the degree of overrejection seemed to be related to how close ρ was to 1. At our baseline parameter value, a nominal 5% test based on Qu's approach had size of approximately 8%, but if we increased ρ to 0.9 or 0.95, we obtained size of approximately 15% and 33%, respectively. While the tests proposed in this paper are not robust to unit roots, they did not show similar sensitivity to the choice of ρ and had roughly the same size for a wide range of values for ρ. Qu suggested that the size distortions of the frequency-domain LM test were due to bias in the periodogram, and proposed a prewhitened version of his test that resolves these size issues in our context.[8] Our power simulations focus on this prewhitened (PW) test, which we call Qu's PW LM test.

TABLE 5. Simulated test size for the full parameter vector (number of simulations is 1000).

Level	$\mathrm{LM}_o(\theta_0)$	$\mathrm{LM}_e(\theta_0)$	Qu LM	Qu PW LM	GQIK	DKK
5%	9%	4.5%	8.4%	4.8%	6.7%	6.4%
10%	15.3%	9.1%	13.6%	8.6%	11.8%	11.5%

[7] Qu's test allows one to test hypotheses using only a subset of frequencies, if desired. For comparability with the other tests studied, we focus on results obtained using the whole spectrum.

[8] In private correspondence with the authors. The prewhitening procedure consists of simulating a long sample under the null and fitting a VAR(1) model to this simulated data. Letting A be the matrix of VAR coefficients and X_T be the $T \times 3$ matrix of data, one then applies Qu's approach using the transformed data $Y_T = X_T(\mathrm{Id}_3 - A \cdot L)$, where L denotes the lag operator. Correspondingly, in all later expressions, the spectral density $f_\theta(\omega)$ is replaced by $g_\theta(\omega) = (\mathrm{Id}_3 - A' \cdot \exp(-i\omega))f_\theta(\omega)(\mathrm{Id}_3 - A' \cdot \exp(-i\omega))^*$, where M^* denotes the conjugate transpose of M.

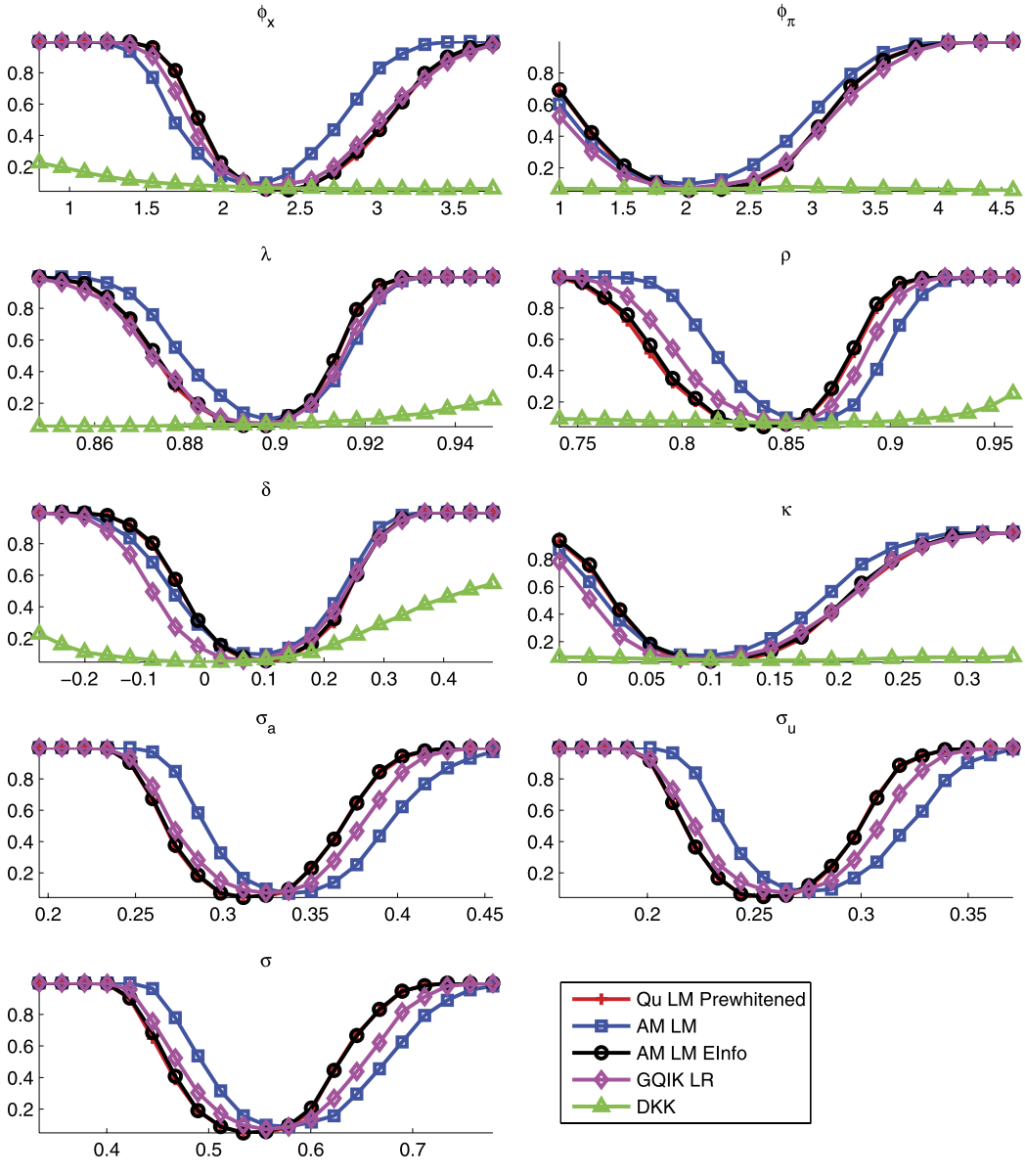

FIGURE 2. Power functions for 5% tests of the null hypothesis $H_0 : \theta = \theta_0$ for the following statistics: $\mathrm{LM}_e(\theta_0)$, $\mathrm{LM}_o(\theta_0)$, prewhitened version of Qu's test, GQIK LR, and DKK test with Newey–West covariance matrix. Power is calculated based on 500 simulations.

We find that the power function for Qu's PW LM test is nearly indistinguishable from the power function for the LM statistic $\mathrm{LM}_e(\theta_0)$ based on theoretical Fisher information. On the one hand, this may seem surprising, since the non-prewhitened version of Qu's test had behavior (in particular, size) that differed substantially from that of the LM_e test. On the other hand, Qu's statistic has the same form as $\mathrm{LM}_e(\theta_0)$ but is calculated with an

approximate likelihood while $\mathrm{LM}_e(\theta_0)$ is calculated with the exact likelihood. The discrepancy between Qu's original LM test and the LM_e test is thus due to the difference between the approximate likelihood and the true likelihood. Insofar as the quasi-likelihood based on the prewhitened data offers a better approximation to the true likelihood, one would expect the behavior of the prewhitened LM test to be closer to that of LM_e. Consistent with this interpretation, the correlation between the prewhitened version of Qu's statistic and $\mathrm{LM}_e(\theta_0)$ under the null is 0.9.

In the GQIK LR approach, rather than testing a hypothesis about the nine-dimensional structural parameter $H_0 : \theta = \theta_0$, one instead tests a hypothesis about the reduced-form parameter (i.e., the coefficients of the state-space representation) $\widetilde{H}_0 : Y = Y(\theta_0)$ using the LR statistic. While simulating GQIK's method, we encountered several difficulties. First, it is not obvious how many degrees of freedom to use. Examining the solution of the model, we noticed that matrices of the state-space representation have numerous zeros. We imposed these zeros, which left us with 28 nonzero reduced-form parameters. However, the effective dimensionality of the reduced-form parameter space is lower since some values of the reduced parameters are observationally equivalent. Hence, we used degrees of freedom equal to the rank of the Fisher information with respect to the state-space coefficients evaluated under the null, which leads us to think that the (local) dimensionality of the reduced-form parameter space is 18.

The second difficulty is that computing the GQIK LR statistic is numerically very involved and time consuming, as noted by GQIK in their paper. To test a hypothesis on the full parameter vector, one must solve a high-dimensional nonlinear optimization problem, while no optimization is required for the other methods discussed here. From Figure 2, one can see that the GQIK test gives us power comparable to the LM tests for all considered alternatives.

For the test of DKK, we consider the transformation of the data

$$\begin{aligned}
\xi_{\pi,t} &= b\pi_{t+1} + \kappa x_t - \pi_t,\\
\xi_{x,t} &= \tilde{\xi}_{x,t} - \rho\tilde{\xi}_{x,t-1},\\
\xi_{r,t} &= \tilde{\xi}_{r,t} - \delta\tilde{\xi}_{r,t-1},
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\xi}_{x,t} &= -[r_t - \pi_{t+1}] + x_{t+1} - x_t;\\
\tilde{\xi}_{r,t} &= \lambda r_{t-1} + (1-\lambda)\phi_\pi \pi_t + (1-\lambda)\phi_x x_t - r_t.
\end{aligned}$$

The transformed data $(\xi_{\pi,t}, \xi_{x,t}, \xi_{r,t})$ comprise a linear combination of the uncorrelated structural error terms $(\varepsilon_t, \varepsilon_{a,t}, \varepsilon_{u,t})$ and the expectation errors $E_t\pi_{t+1} - \pi_{t+1}$, $E_{t-1}\pi_t - \pi_t$, $E_t x_{t+1} - x_{t+1}$, and $E_{t-1}x_t - x_t$. We base the test on the exclusion restriction that $(\xi_{\pi,t}, \xi_{x,t}, \xi_{r,t})$ are not predictable by the instruments $Y_{t-1} = (\pi_{t-1}, x_{t-1}, r_{t-1})$. It is easy to see that $(\xi_{\pi,t}, \xi_{x,t}, \xi_{r,t})$ follows a (moving average) MA(1) process and hence that the heteroskedasticity and autocorrelation robust (HAC) formulation of DKK should be used. We calculate the DKK test using the Newey–West HAC estimator for the long-run covariance matrix (using three lags). DKK formulate the null in such a way that variances

of the shocks do not enter, and the test is not supposed to have power against alternatives that differ only in these parameters. Hence, we do not depict the corresponding power functions.

Based on Figure 2, the DKK test in our context is significantly less powerful than the other tests considered and has nearly flat power curves in the neighborhoods where the LM tests achieve almost 100% power. Power simulations on larger neighborhoods show that the DKK test has nontrivial power against some alternatives, but confirm that for the null and alternatives considered, it has substantially less power than the other tests we study. This lower power is to be expected given the limited-information nature of the test, and may be a reasonable price to pay for robustness to misspecification.

8. Conclusion

This paper studies the problem of weak identification in DSGE models and explores how weak identification can arise in several examples. We show that two forms of the LM statistic may be used to construct robust tests for hypotheses about the full parameter vector, as well as hypotheses about subvectors of parameters for which the nuisance parameter is strongly identified. How to determine whether the nuisance parameter is strongly identified is an open question. We give suggestive evidence that the discrepancy between two measures of information may serve as an indication of weak identification, but further exploration of this issue is an important topic for future research.

Appendix: Proofs

We denote by superscript 0 quantities evaluated at $\theta_0 = (\alpha_0', \beta_0')'$. In the Taylor expansions used in the proofs, 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

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

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 central limit theorem (CLT) for $K_{\alpha,T}S_{\alpha,T}$. □

Lemma 2. *Let $M_T = \sum_{t=1}^T m_t$ be a multidimensional 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 ith component of m_t and let $K_{i,T}$ be the ith diagonal element of K_T. For any i, j, l,*

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

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

$$\big(K_{\alpha,T}S^0_\alpha, K_{\beta,T}S^0_\beta, K_{\alpha\beta,T}\operatorname{vec}\big(A^0_{\alpha\beta}\big)'\big) \Rightarrow (\xi_\alpha, \xi_\beta, \xi_{\alpha\beta}), \tag{8}$$

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

We Taylor expand $S_{\beta_j}(\hat\alpha, \beta_0)$, the jth 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

$$\begin{aligned} K_{\beta_j,T}S_{\beta_j}(\hat\alpha, \beta_0) = {} & K_{\beta_j,T}S^0_{\beta_j} - K_{\beta_j,T}I^0_{\beta_j\alpha}(\hat\alpha - \alpha_0) \\ & + \frac{1}{2}K_{\beta_j,T}(\hat\alpha - \alpha_0)'\big(I^0_{\alpha\alpha\beta_j}\big)(\hat\alpha - \alpha_0) + \tilde R_j \end{aligned}$$

with residual

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

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 a point between $\hat\alpha$ and α_0. From Assumption 2(c), 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:

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

We plug asymptotic statement (7) into this equation and get

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

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

$$\begin{aligned} K_{\beta_j,T}S_{\beta_j}(\hat\alpha, \beta_0) = {} & K_{\beta_j,T}S^0_{\beta_j} - K_{\beta_j,T}J^0_{\beta_j\alpha}\big(I^0_{\alpha\alpha}\big)^{-1}S^0_\alpha + K_{\beta_j,T}A^0_{\beta_j\alpha}\big(I^0_{\alpha\alpha}\big)^{-1}S^0_\alpha \\ & + \frac{1}{2}K_{\beta_j,T}S^{0\prime}_\alpha\big(I^0_{\alpha\alpha}\big)^{-1}\big(I^0_{\alpha\alpha\beta_j}\big)\big(I^0_{\alpha\alpha}\big)^{-1}S^0_\alpha + o_p(1). \end{aligned} \tag{9}$$

One can notice that we have the informational equality

$$I^0_{\alpha\alpha\beta_j} = -\big[A^0_{\alpha\alpha}, S^0_{\beta_j}\big] - \big[A^0_{\alpha\beta_j}, S^0_\alpha\big] - \big[S^0_\alpha, A^0_{\alpha\beta_j}\big] + 2\sum_{t=1}^{T} s_{\alpha,t}s'_{\alpha,t}s_{\beta_j,t} + \Lambda_{\alpha\alpha\beta_j}. \tag{10}$$

Assumption 4(b) implies that $K_{\beta_j,T}K_{\alpha,T}\Lambda_{\alpha\alpha\beta_j}K_{\alpha,T} \xrightarrow{p} 0$. Assumption 2(a) and Assumption 3 together imply that $(K_{\alpha,T}\otimes K_{\alpha,T})K_{\alpha\alpha,T}^{-1} \to 0$. Using Assumption 2(a) and Lemma 2, we notice that

$$\begin{aligned} -K_{\alpha,T}I^0_{\alpha\alpha\beta_j}K_{\alpha,T} &= K_{\alpha,T}\big[A^0_{\alpha\beta_j}, S^0_\alpha\big]K_{\alpha,T} + K_{\alpha,T}\big[S^0_\alpha, A^0_{\alpha\beta_j}\big]K_{\alpha,T} \\ &\quad + o_p\big(K^{-1}_{\beta_j,T}\big). \end{aligned} \tag{11}$$

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 (9). Doing so gives us

$$\begin{aligned} K_{\beta_j,T}S_{\beta_j}(\hat\alpha,\beta_0) &= K_{\beta_j,T}S^0_{\beta_j} - K_{\beta_j,T}J^0_{\beta_j\alpha}\big(J^0_{\alpha\alpha}\big)^{-1}S^0_\alpha + K_{\beta_j,T}A^0_{\beta_j\alpha}\big(J^0_{\alpha\alpha}\big)^{-1}S^0_\alpha \\ &\quad + \frac{1}{2}K_{\beta_j,T}S^{0\prime}_\alpha\big(J^0_{\alpha\alpha}\big)^{-1}\big(I^0_{\alpha\alpha\beta_j}\big)\big(J^0_{\alpha\alpha}\big)^{-1}S^0_\alpha + o_p(1), \\ K_{\beta_j,T}S_{\beta_j}(\hat\alpha,\beta_0) &= K_{\beta_j,T}S^0_{\beta_j} - K_{\beta_j,T}J^0_{\beta_j\alpha}\big(J^0_{\alpha\alpha}\big)^{-1}S^0_\alpha + D'_j\big(J^0_{\alpha\alpha}K_{\alpha,T}\big)^{-1}S^0_\alpha + o_p(1), \end{aligned} \tag{12}$$

where

$$D_j = K_{\alpha,T}K_{\beta_j,T}A^0_{\alpha\beta_j} + \frac{1}{2}K_{\alpha,T}K_{\beta_j,T}\big(I^0_{\alpha\alpha\beta_j}\big)\big(J^0_{\alpha\alpha}\big)^{-1}S^{0\prime}_\alpha.$$

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 (11) we have

$$\begin{aligned} D_j \;=\; & K_{\alpha,T}K_{\beta_j,T}K^{-1}_{\alpha\beta_j,T} \\ & \times \big(K_{\alpha\beta_j,T}A^0_{\alpha\beta_j} - \big(K_{\alpha\beta_j,T}\big[A^0_{\alpha\beta_j}, S^0_\alpha\big]K_{\alpha,T}\big)\big(K_{\alpha,T}J^0_{\alpha\alpha}K_{\alpha,T}\big)^{-1}K_{\alpha,T}S^{0\prime}_\alpha\big) \\ & + o_p(1) \\ \Rightarrow\; & C\big(\xi_{\alpha\beta_j} - \operatorname{cov}(\xi_{\alpha\beta_j},\xi_\alpha)\operatorname{Var}(\xi_\alpha)^{-1}\xi_\alpha\big), \end{aligned}$$

where variables $(\xi'_\alpha, \xi'_{\alpha\beta_j})$ are as described at the beginning of the proof.

Plugging the last statement and (8) into equation (12), we have

$$\begin{aligned} K_{\beta_j,T}S_{\beta_j}(\hat\alpha,\beta_0) \Rightarrow\; & \xi_{\beta_j} - \operatorname{cov}(\xi_{\beta_j},\xi_\alpha)\operatorname{Var}(\xi_\alpha)^{-1}\xi_\alpha \\ & + C\big(\xi_{\alpha\beta_j} - \operatorname{cov}(\xi_{\alpha\beta_j},\xi_\alpha)\operatorname{Var}(\xi_\alpha)^{-1}\xi_\alpha\big)\operatorname{Var}(\xi_\alpha)^{-1}\xi_\alpha. \end{aligned} \tag{13}$$

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 (6) for $\widetilde{\mathrm{LM}}_o(\beta_0)$, which is equal to $(J_{\beta\beta} - J_{\beta\alpha}J^{-1}_{\alpha\alpha}J'_{\beta\alpha})|_{(\hat\alpha,\beta_0)}$. Below we prove the following lemma.

LEMMA 3. *Under the assumptions of Theorem 2, we have*

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

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

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

Lemma 3 implies that

$$\begin{aligned}
&K_{\beta_i,T}K_{\beta_j,T}\big(J_{\beta_i\beta_j} - J_{\beta_i\alpha}J_{\alpha\alpha}^{-1}J_{\beta_j\alpha}'\big)\big|_{(\hat{\alpha},\beta_0)}\\
&\quad\Rightarrow \operatorname{cov}(\xi_{\beta_i},\xi_{\beta_j}) + C\cdot\operatorname{cov}(\xi_{\alpha\beta_i},\xi_{\beta_j})'\operatorname{Var}(\xi_\alpha)^{-1}\xi_\alpha\\
&\qquad + C\cdot\operatorname{cov}(\xi_{\alpha\beta_j},\xi_{\beta_i})'\operatorname{Var}(\xi_\alpha)^{-1}\xi_\alpha\\
&\qquad + C^2\xi_\alpha'\operatorname{Var}(\xi_\alpha)^{-1}\operatorname{cov}(\xi_{\alpha\beta_i},\xi_{\alpha\beta_j})\operatorname{Var}(\xi_\alpha)^{-1}\xi_\alpha\\
&\qquad - \big(\operatorname{cov}(\xi_\alpha,\xi_{\beta_i}) + C\cdot\operatorname{cov}(\xi_{\alpha\beta_i},\xi_\alpha)\operatorname{Var}(\xi_\alpha)^{-1}\xi_\alpha\big)'\operatorname{Var}(\xi_\alpha)^{-1}\\
&\qquad \times \big(\operatorname{cov}(\xi_\alpha,\xi_{\beta_j}) + C\cdot\operatorname{cov}(\xi_{\alpha\beta_j},\xi_\alpha)\operatorname{Var}(\xi_\alpha)^{-1}\xi_\alpha\big).
\end{aligned}$$

Note that the last expression is the same as the variance of the right side of equation (13) conditional on random variable ξ_α. That is, $K_{\beta,T}(J_{\beta\beta} - J_{\beta\alpha}J_{\alpha\alpha}^{-1}J_{\beta\alpha}')K_{\beta,T}|_{(\hat{\alpha},\beta_0)}$ 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 $\widetilde{\mathrm{LM}}_o(\beta_0)$, conditional on ξ_α, is distributed $\chi^2_{k_\beta}$ asymptotically and thus is asymptotically $\chi^2_{k_\beta}$ unconditionally as well. The case of statistic $\widetilde{\mathrm{LM}}_e(\beta_0)$ is analogous. This completes the proof of Theorem 2. □

PROOF OF LEMMA 3. (a) We can Taylor expand $J_{\beta_i\beta_j}(\hat{\alpha},\beta_0)$ as

$$\begin{aligned}
J_{\beta_i\beta_j}(\hat{\alpha},\beta_0) = J^0_{\beta_i\beta_j} + \frac{\partial}{\partial\alpha}J^0_{\beta_i\beta_j}(\hat{\alpha}-\alpha_0)\\
+ \frac{1}{2}(\hat{\alpha}-\alpha_0)'\frac{\partial^2}{\partial\alpha\,\partial\alpha'}J^0_{\beta_i\beta_j}(\hat{\alpha}-\alpha_0) + R_{ij},
\end{aligned} \tag{14}$$

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 2 and Assumption 4(a), we have

$$K_{\alpha,T}K_{\beta_i,T}K_{\beta_j,T}\frac{\partial}{\partial\alpha'}J_{\beta_i\beta_j} \xrightarrow{p} C\cdot\operatorname{cov}(\xi_{\alpha\beta_i},\xi_{\beta_j}) + C\cdot\operatorname{cov}(\xi_{\alpha\beta_j},\xi_{\beta_i}). \tag{15}$$

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

$$\begin{aligned}
&K_{\beta_i,T}K_{\beta_j,T}K_{\alpha,T}\frac{\partial^2}{\partial\alpha\,\partial\alpha'}J_{\beta_i\beta_j}K_{\alpha,T}\\
&\quad= K_{\beta_i,T}K_{\beta_j,T}K_{\alpha,T}\\
&\qquad\times\big([\Lambda_{\alpha\alpha\beta_i},S_{\beta_j}]+[\Lambda_{\alpha\alpha\beta_j},S_{\beta_i}]+[A_{\alpha\beta_i},A_{\alpha\beta_j}]+[A_{\alpha\beta_j},A_{\alpha\beta_i}]\big)K_{\alpha,T}+o_p(1).
\end{aligned}$$

The $o_p(1)$ term appears due to Lemma 2, applied to the remaining terms. Assumption 4(b) implies that $K_{\alpha,T}K_{\beta_i,T}K_{\beta_j,T}[\Lambda_{\alpha\alpha\beta_i},S_{\beta_j}]K_{\alpha,T}\xrightarrow{p}0$. Finally, using Assumption 3(b), we get

$$\begin{aligned}
&K_{\beta_i,T}K_{\beta_j,T}K_{\alpha,T}\frac{\partial^2}{\partial\alpha\,\partial\alpha'}J_{\beta_i\beta_j}K_{\alpha,T}\\
&\quad\xrightarrow{p}C^2\operatorname{cov}(\xi_{\alpha\beta_i},\xi_{\alpha\beta_j})+C^2\operatorname{cov}(\xi_{\alpha\beta_j},\xi_{\alpha\beta_i}).
\end{aligned}\tag{16}$$

Putting the expressions for derivatives (15) and (16) into equation (14), and also noticing that due to Lemma 1, $K_{\alpha,T}^{-1}(\hat\alpha-\alpha_0)\Rightarrow\operatorname{Var}(\xi_\alpha)^{-1}\xi_\alpha$, we get statement (a).

(b) Again we use Taylor expansion:

$$\begin{aligned}
J_{\alpha\beta_j}(\hat\alpha,\beta_0)&=J^0_{\alpha\beta_j}+\frac{\partial}{\partial\alpha}J^0_{\alpha\beta_j}(\hat\alpha-\alpha_0)\\
&\quad+\frac12\sum_n\frac{\partial^2}{\partial\alpha\,\partial\alpha_n}J^*_{\alpha\beta_j}(\hat\alpha-\alpha_0)(\hat\alpha_n-\alpha_{0,n}).
\end{aligned}\tag{17}$$

From Assumption 3(b),

$$K_{\alpha,T}K_{\beta_j,T}J^0_{\alpha\beta_j}\xrightarrow{p}\operatorname{cov}(\xi_\alpha,\xi_{\beta_j}).\tag{18}$$

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 2, $K_{\alpha,T}K_{\beta_j,T}\sum s_{\alpha,t}s'_{\alpha,t}s_{\beta_j,t}K_{\alpha,T}\to0$. Assumptions 2(a) and 3 imply that $K_{\alpha,T}K_{\beta_j,T}[A_{\alpha\alpha},S_{\beta_j}]K_{\alpha,T}\xrightarrow{p}0$. We have

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

Similarly, we can show that the residual term in (17) is asymptotically negligible. Putting the last equation, together with (18), into (17) and using Lemma 1, we get statement (b) of Lemma 3.

(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, $K_{\alpha,T}K_{\alpha_n,T}[A_{\alpha\alpha_n}, S_\alpha]K_{\alpha,T} \xrightarrow{p} 0$, and according to Lemma 2, $K_{\alpha,T}K_{\alpha_n,T}\sum s_{\alpha,t}s'_{\alpha,t}s_{\alpha_n,t}K_{\alpha,T} \xrightarrow{p} 0$. Given the result of Lemma 1, we arrive at statement (c). □

References

Altug, S. (1989), "Time-to-build aggregate fluctuations: Some new evidence." *International Economic Review*, 30 (4), 889–920. [123]

Amisano, G. and O. Tristani (2011), "Exact likelihood computation for nonlinear DSGE models with heteroskedastic innovations." *Journal of Economic Dynamics & Control*, 35 (12), 2167–2185. [141]

Andrews, D. W. K. (1987), "Asymptotic results for generalized Wald tests." *Econometric Theory*, 3, 348–358. [132]

Andrews, D. W. K. and X. Cheng (2012), "Estimation and inference with weak, semi-strong and strong identification." *Econometrica*, 80 (5), 2153–2211. [127, 130]

Barndorff-Nielsen, O. E. and M. Sorensen (1991), "Information quantities in non-classical settings." *Computational Statistics & Data Analysis*, 12, 143–158. [129]

Basawa, I. and H. L. Koul (1979), "Asymptotic tests of composite hypotheses for non-ergodic type stochastic processes." *Stochastic Processes and Their Applications*, 9, 291–305. [131]

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. [133, 145]

Canova, F. and L. Sala (2009), "Back to square one: Identification issues in DSGE models." *Journal of Monetary Economics*, 56, 431–449. [124, 125]

Clarida, R., J. Gali, and M. Gertler (1999), "The science of monetary policy: A new Keynesian perspective." *Journal of Economic Literature*, 37, 1661–1707. [136]

Dufour, J. M., L. Khalaf, and M. Kichian (2013), "Identification-robust analysis of DSGE and structural macroeconomic models." *Journal of Monetary Economics*, 60 (3), 340–350. [123, 125, 140]

Dufour, J. M. and M. Taamouti (2005), "Projection-based statistical inference in linear structural models with possibly weak instruments." *Econometrica*, 73, 1351–1365. [132]

Fernández-Villaverde, J. (2010), "The econometrics of DSGE models." *SERIEs: Journal of the Spanish Economic Association*, 1, 3–49. [123]

Fernández-Villaverde, J. and J. F. Rubio-Ramírez (2011), "Macroeconomics and volatility: Data, models, and estimation." In *Advances in Economics and Econometrics: Theory and Applications, Tenth World Congress* (D. Acemoglu, M. Arellano, and E. Dekel, eds.), Cambridge University Press, Cambridge. [141]

Guerron-Quintana, P., A. Inoue, and L. Kilian (2013), "Frequentist inference in weakly identified DSGE models." *Quantitative Economics*, 4 (2), 197–229. [124, 125, 140]

Hall, P. and C. C. Heyde (1980), *Martingale Limit Theory and Its Application*. Academic Press, New York. [129]

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. [123]

Ireland, P. N. (2004), "Technology shocks in the new Keynesian model." *Review of Economics and Statistics*, 86, 923–936. [123]

Iskrev, N. (2008), "Evaluating the information matrix in linearized DSGE models." *Economics Letters*, 99, 607–610. [135]

Iskrev, N. (2010), "Evaluating the strength of identification in DSGE models. An a priori approach." Working paper, Bank of Portugal. [124, 132, 135]

Jerez, M., J. Casals, and S. Sotoca (2011), *Signal Extraction for Linear State Space Models*, Lambert Academic Publishing, Saarbrucken. [135]

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. [133]

Komunjer, I. and S. Ng (2011), "Dynamic identification of dynamic stochastic general equilibrium models." *Econometrica*, 79 (6), 1995–2032. [132]

Lindé, J. (2005), "Estimating new-Keynesian Phillips curves: A full information maximum likelihood approach." *Journal of Monetary Economics*, 52 (6), 1135–1149. [123]

Liptser, R. and A. Shiryayev (1989), *Theory of Martingales*. Springer, Berlin. [131]

Mavroeidis, S. (2005), "Identification issues in forward-looking models estimated by GMM with an application to the Phillips curve." *Journal of Money, Credit, and Banking*, 37, 421–449. [124, 125]

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. [123]

Qu, Z. (forthcoming), "Inference and specification testing in DSGE models with possible weak identification." *Quantitative Economics*. [125, 140, 141, 142]

Schorfheide, F. (2010), "Estimation and evaluation of DSGE models: Progress and challenges." Working paper, NBER. [141]

Silvey, S. D. (1961), "A note on maximum-likelihood in the case of dependent random variables." *Journal of the Royal Statistical Society, Series B, Methodological*, 23, 444–452. [128]

Smets, F. and R. Wouters (2007), "Shocks and frictions in US business cycles: A Bayesian DSGE approach." *American Economic Review*, 97 (3), 586–606. [137]

Stock, J. H. and J. H. Wright (2000), "GMM with weak identification." *Econometrica*, 68, 1055–1096. [127, 133, 141]

White, H. (1982), "Maximum likelihood estimation in misspecified models." *Econometrica*, 50, 1–25. [124, 130]

Submitted November, 2012. Final version accepted December, 2013.