Weak Identification in Maximum Likelihood: A Question of Information

By Isaiah Andrews and Anna Mikusheva*

Weak identification commonly refers to the failure of classical asymptotics to provide a good approximation to the finite sample distribution of estimates and t- and Wald statistics in point-identified models where the data contains little information. There are several commonly accepted ways of modeling this situation, which include the drifting objective function approach of Stock and Wright (2000) and the drifting parameter approach used in D. Andrews and Cheng (2012). Unfortunately there are empirically relevant contexts, for example many Dynamic Stochastic General Equilibrium (DSGE) models, where simulation evidence strongly suggests weak identification but it is unclear how to cast the model into either of these frameworks. Concerns about weak identification in DSGE models were raised in a number of papers (see for example Canova and Sala (2009) and Schorfheide (2013)). At the same time, due in part to the analytical intractability of these models, the sources and nature of weak identification and the routes through which weak identification distorts non-robust approaches to inference are not yet clear.

Here we highlight a previously overlooked feature common to many weakly identified models which plays an important role in the behavior of the maximum likelihood estimator (MLE). The usual approximations for the MLE rely critically on the assumption that two approaches to estimating Fisher information, through the quadratic variation of the score and the negative Hessian of the log-likelihood, provide nearly identical answers. We show that in many weakly identified contexts the appropriately normalized quadratic variation of the score converges to the normalized Fisher information, but that the normalized negative Hessian remains volatile even in large samples. To capture this effect, we introduce a measure of the disparity between the two estimators of information, which will converge to zero in strongly identified contexts but can otherwise distort the distribution of the MLE. Using simulations in a stylized DSGE model we show that this discrepancy between information measures becomes large precisely when the classical asymptotic approximations are especially unreliable.

This paper is closely related to Andrews and Mikusheva (2013) (henceforth AM), where we provide additional examples and discuss tests which are insensitive to the disparity between the two estimates of information and are robust to weak identification.

I. Likelihood Theory

Let $X_T = (x_1, ..., x_T)$ be the data available at time T, and let $\mathcal{F}_T$ be the sigma-algebra generated by X_T. We consider parametric models where the log-likelihood $\ell(X_T; \theta) = \log f(X_T; \theta)$ is known up to the k-dimensional parameter θ which has true value θ_0. We further assume that $\ell(X_T; \theta)$ is twice continuously differentiable with respect to θ. If we have correctly specified the model, the score $S_T(\theta) = \frac{\partial}{\partial \theta'} \ell(X_T, \theta)$, evaluated at the true parameter value θ_0, is a martingale with respect to filtration $\mathcal{F}_t$ under mild conditions.

We consider two measures of information based on observed quantities. The first

* Andrews: Department of Economics, M.I.T., 77 Massachusetts Avenue, E19-750, Cambridge, MA-02142, iandrews@mit.edu. Mikusheva: Department of Economics, M.I.T., 77 Massachusetts Avenue, E18-224, Cambridge, MA-02142, amikushe@mit.edu. Andrews acknowledges financial support from the NSF Graduate Research Fellowship under Grant No. 1122374. Mikusheva acknowledges financial support from the Castle-Krob Career Development Chair and Sloan Research Fellowship.

one, *observed information*, equals the negative Hessian of the log-likelihood $I_T(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta'}\ell(X_T;\theta)$. The second, *incremental observed information,* equals the quadratic variation of the score,

$$J_T(\theta) = [S(\theta)]_T = \sum_{t=1}^{T} s_t(\theta)s_t'(\theta),$$

where $s_t(\theta) = S_t(\theta) - S_{t-1}(\theta)$. In what follows we will take I_T and J_T, written without arguments, to denote $I_T(\theta_0)$ and $J_T(\theta_0)$. If the model is correctly specified both I_T and J_T may serve as estimates of the (theoretical) Fisher information for the whole sample, and by the second informational equality $E(I_T) = E(J_T)$.

In the classical context I_T and J_T are asymptotically equivalent, which plays a key role in the asymptotics of maximum likelihood. The asymptotic normality of the MLE is driven by two key assumptions: (i) that the log-likelihood is asymptotically locally quadratic and (ii) that the difference between the two measures of information I_T and J_T is small asymptotically (see Geyer (2013)). Specifically, using the first order condition for likelihood maximization one can show that for $\hat\theta$ the MLE,

$$\begin{aligned} J_T^{1/2}(\hat\theta - \theta_0) &= J_T^{-1/2}S_T(\theta_0) + \\ &+ J_T^{-1/2}(I_T - I_T(\theta^*))J_T^{-1/2}J_T^{1/2}(\hat\theta - \theta_0) + \\ &+ J_T^{-1/2}(J_T - I_T)J_T^{-1/2}J_T^{1/2}(\hat\theta - \theta_0), \end{aligned} \tag{1}$$

where θ^* is a point in between $\hat\theta$ and θ_0 which may differ across rows of $I_T(\theta^*)$. The first term, $J_T^{-1/2}S_T(\theta_0)$, is asymptotically standard normal under quite general conditions as discussed in AM. Provided $J_T^{1/2}(\hat\theta - \theta_0)$ is stochastically bounded the second term in (1) is small so long as the log-likelihood is close to quadratic on a neighborhood containing both θ_0 and $\hat\theta$. In this paper we will focus on the third term in (1), and in particular on the standardized difference between information measures $J_T^{-1/2}(J_T - I_T)J_T^{-1/2}$, which can render the usual asymptotic approximations to the behavior of the MLE quite poor if it is large. We argue that in weakly identified models the difference between the two observed measures of information may not be negligible compared to observed incremental information J_T and that the third term in (1) thus plays an important role in the behavior of the MLE under weak identification.

II. Two Estimates of Information

Here we highlight the importance of the standardized difference between information measures, $J_T^{-1/2}(J_T - I_T)J_T^{-1/2}$, under weak identification. We begin by noting that this term is asymptotically non-trivial in a number of weakly identified examples, including a simple linear instrumental variables model.

Example. Consider a homoskedastic linear instrumental variables model

$$\begin{cases} Y = \beta Z\pi + U \\ X = Z\pi + V \end{cases},$$

where Y and X are endogenous variables while Z is a $T \times k$ matrix of exogenous instruments. We assume that $Z'Z/T$ converges in probability to Q and $Z'[U,V]/\sqrt{T}$ converges in distribution to $N(0, \Sigma \otimes Q)$ as the sample size T increases, for Q a full rank matrix and Σ the covariance matrix of the reduced form errors. We consider a Gaussian likelihood as a function of the structural parameters $\theta = (\pi', \beta)'$. Weak instruments are usually modeled by considering a sequence of models in which the correlation between the instruments and the endogenous regressor drifts towards zero as the sample size increases, $\pi = \pi_T = c/\sqrt{T}$, with the consequence that information about the value of β does not increase with the sample size. Under such weak sequences, for a $(k+1) \times (k+1)$ normalization matrix $K_T = diag(1/\sqrt{T}, ..., 1/\sqrt{T}, 1)$, $K_T J_T K_T$ converges in probability to a non-random positive definite matrix $\mathcal{J}$ while $K_T I_T K_T$ converges in distribution to a random Gaussian matrix with mean $\mathcal{J}$. To characterize the asymptotic disparity between the two estimators of the Fisher information we can consider $M = J_T^{-1/2}(I_T - J_T)J_T^{-1/2}$. Under weak instrument asymptotics the trace

of M converges in distribution to a mean zero Gaussian random variable with variance equal to the inverse of the concentration parameter (which measures the informativeness of the instruments, see Staiger and Stock (1997)) multiplied by a measure of the degree of endogeneity. In particular, when the instruments are nearly irrelevant M will be (stochastically) large.

This asymptotic disparity between the two estimates of the Fisher information also appears in a number of other weakly identified models. In AM we showed that this issue arises in an ARMA(1,1) model with nearly canceling roots, VAR models with weakly identified dynamics, weakly identified exponential family models, and weakly identified mixture models. In all of these models, J_T is positive-definite with probability one and, appropriately normalized, converges in probability to a non-random positive definite matrix. If one applies the same normalization to I_T then in the strongly identified case it converges to the same limit as J_T but in the weakly identified case it converges in distribution to a random matrix. This random matrix has mean equal to the limit of the normalized J_T, as suggested by the second informational equality, but has non-trivial variance.

We emphasize four important points. First, the question of how to define, model, and measure weak identification is still open in many contexts. There are some models, like homoskedastic weak IV, in which we know how to directly measure identification strength (the concentration parameter). There are other models, like those studied by Stock and Wright (2000), where we have theoretical approaches to model weak identification but have no way to measure whether weak identification is a problem in a given empirical application. Finally there are many contexts, like DSGE models (see Canova and Sala (2009)), in which we strongly suspect that weak identification is a problem but still largely lack tools to model or measure it. We suggest that the size of matrix

$$M = J_T^{-1/2} \left(I_T - J_T \right) J_T^{-1/2}$$

is an important reflection of identification strength in parametric models. As already discussed M is asymptotically nontrivial in a number of weakly identified examples and, as we can see from expansion (1), large values of M can introduce distortions in the classical MLE asymptotics.

Second, while it is common to associate weak identification with the Fisher information $EJ_T = EI_T$ being nearly degenerate or the likelihood being nearly flat along some directions, we argue that these are misleading characterizations as neither the Fisher information nor the Hessian of the likelihood are invariant to re-parametrization. In particular, if we linearly re-parameterize a model in terms of $\tau = \frac{\theta}{k}$ then both measures of information scale by a factor k^2. Hence, by linear re-parametrization one can produce a model whose Fisher information is arbitrarily small (or large) without changing the quality of the classical ML approximation. Consequently, any approach which detects weak identification by assessing how close the information is to degeneracy, for example Iskrev (2010), is misleading. In our examples weak identification is associated with the curvature of the objective function (the negative Hessian I_T) being different from J_T even in very large samples, so we think it is potentially more fruitful to associate weak identification with a low signal-to-noise ratio, treating J_T as the signal and $I_T - J_T$ as noise, suggesting the measure $M = J_T^{-1/2} \left(I_T - J_T \right) J_T^{-1/2}$.

Third, this disparity between two estimates of the Fisher information is not a sign of mis-specification, as even in correctly specified models these two measures may differ substantially if identification is weak. Correct specification implies that $EJ_T = EI_T$, and it is this restriction that is tested by the Information Matrix Test of White (1982). In contrast, weak identification is related to $I_T - J_T$ being volatile relative to J_T, but the restriction $EJ_T = EI_T$ continues to hold under correct specification.

Fourth, the classical asymptotic approximations for the MLE and Wald statistic require that the disparity measure M be small. By contrast, the distribution of the

robust score (LM) tests discussed in AM is insensitive to the behavior of M, and these tests remain well-behaved in weakly identified settings.

III. A Small DSGE Model

In this section we examine the effects of weak identification on estimation and inference in a simple DSGE model. Most DSGE models must be solved numerically, and it is typically difficult to say which parameters are weakly identified and what aspects of the model give rise to weak identification. To overcome these difficulties, here we study a highly stylized DSGE model which can be solved analytically, allowing us to explicitly model weak identification.

Assume we observe inflation π_t and a measure of real activity x_t which obey

$$\begin{cases} bE_t\pi_{t+1} + \kappa x_t - \pi_t = 0, \\ r_t - E_t\pi_{t+1} - \rho\Delta a_t = E_t x_{t+1} - x_t, \\ \frac{1}{b}\pi_t + u_t = r_t. \end{cases}$$

where E_t denotes $E[\cdot|\mathcal{F}_t]$. The first equation is a linearized Euler equation while the second is a Phillips curve. We assume that the interest rate r_t is unobserved, and that the exogenous shocks Δa_t and u_t are generated by:

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

The model has six unknown scalar parameters: the discount fact b, the Calvo parameter κ, the persistence parameters ρ and δ, and the standard deviations σ_a and σ_u. AM show that the model is point identified for $\kappa > 0, \sigma_a^2 > 0, \sigma_u^2 > 0$, and $-1 < \delta < \rho < 1$. By contrast, when $\rho = \delta$ the model is not point identified. We can think of $\rho - \delta$ as controlling identification strength: the model is weakly identified when this difference is small.

To explore the effects of weak identification in this context, we simulate data from the model for different values of $\rho - \delta$. In particular we calibrate the parameters $(b, \kappa, \delta, \sigma_a, \sigma_u)$ to their values in the simulation section of AM, $(.99, .1, .1, .325, .265)$, and consider a range of values for $\rho - \delta$, where for each value of this difference we simulate samples of size 200 from the model. To avoid issues arising from the fact that b is close to its upper bound ($b = 1$), we fix this parameter at its true value and take $\theta = (\kappa, \rho, \delta, \sigma_u, \sigma_v)$ to be the unknown structural parameter. In each sample we calculate the maximum likelihood estimator $\hat{\theta}$, the (non-robust) Wald statistic $(\hat{\theta} - \theta_0)' I(\hat{\theta})(\hat{\theta} - \theta_0)$, and the (robust) score statistic LM_e discussed by AM. The corresponding tests reject when the appropriate statistic exceeds a χ^2_5 critical value. We assess the normality of the MLE by considering the normalized statistic $\mathcal{T} = J_T^{1/2}(\hat{\theta} - \theta_0)$, which converges to a 5-dimensional standard normal vector under strong identification. We calculate the simulation mean and variance of $\mathcal{T}$ and report the deviation of these quantities from zero and the identity matrix, respectively, which should be small if this term is approximately standard normal. Note that while the population mean and variance of $\mathcal{T}$ need not exist, its sample mean and variance in our simulations are always well-defined. Finally, we report some summary statistics for the disparity measure M, in particular the standard deviation of $trace(M)$ and the median of the largest eigenvalue of M in absolute value, both of which should be small if identification is strong. All results are reported in Table 1.

As we can see in Table 1, the standard normal approximation to $\mathcal{T} = J_T^{1/2}(\hat{\theta} - \theta_0)$ breaks down for small values of $\rho - \delta$, as does size control for Wald tests. The behavior of M is similarly sensitive to identification strength, and this term is large precisely when the conventional strong-identification approximations break down. The range of values $\rho - \delta$ which qualify as "small" is surprisingly large: even for $\rho - \delta$ equal to 0.3 the Wald test exhibits substantial size distortions, with rejection probability exceeding 25%. By contrast, the LM_e test is largely insensitive to identification strength. Thus, we can again see that the scaled difference between the two measures of information is (stochastically) large when

TABLE 1—BEHAVIOR OF TESTS AND INFORMATION ESTIMATORS AS A FUNCTION OF $\rho - \delta$ IN DSGE MODEL WITH 200 OBSERVATIONS.

$\rho - \delta$	0.05	0.1	0.2	0.3	0.5	0.7
$\left\|\widehat{E}(\mathcal{T})\right\|$	2,015	309	4.25	1.43	0.57	0.49
$\left\|\widehat{Var}(\mathcal{T}) - Id_5\right\|$	$1.7{\cdot}10^{10}$	$8.5{\cdot}10^{8}$	23.3	3.14	0.43	0.78
$\widehat{Std}(tr(M))$	212	57.8	11.9	3.14	0.85	0.60
Median of $\|M\|$	129	35.4	7.17	2.10	0.82	0.70
Size of 5 percent Wald Test	88.9%	79.8%	52.5%	28.1%	12.1%	9.8%
Size of 5 percent LM_e Test	5.3%	5.4%	5.1%	5.5%	5.2%	5.9%

Note: All quantities based on 10,000 simulation replications, and $\widehat{E}(\cdot)$, $\widehat{Std}(\cdot)$ $\widehat{Var}(\cdot)$ are simulation mean, standard deviation, and variance, respectively. For X a vector $\|X\|$ denotes the Euclidean norm, while for X a square matrix $\|X\|$ denotes the largest eigenvalue of X in absolute value.

identification is weak, and that even in this very simple DSGE model weak identification leads to poor behavior for classical inference procedures over much of the parameter space.

REFERENCES

Andrews, Donald W.K., and Xu Cheng. 2012. "Estimation and Inference with Weak, Semi-strong and Strong Identification." *Econometrica* 80(5): 2153-211.

Andrews, Isaiah, and Anna Mikusheva. 2013. "Maximum Likelihood Inference in Weakly Identified DSGE Models." http://economics.mit.edu/files/6593.

Canova, Fabio, and Luca Sala. 2009. "Back to Square One: Identification Issues in DSGE Models." *Journal of Monetary Economics* 56(4): 431-49.

Geyer, Charles J. 2013. "Asymptotics of Maximum Likelihood without the LLN or CLT or Sample Size Going to Infinity." *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton,* edited by Galin Jones and Xiaotong Shen, 1-24. Beachwood, Ohio, USA: Institute of Mathematical Statistics.

Iskrev, Nikolay. 2010. "Evaluating the Strength of Identification in DSGE Models. An a Priori Approach." Bank of Portugal Working Paper.

Schorfheide, Frank. 2013. "Estimation and Evaluation of DSGE Models: Progress and Challenges." In *Advances in Economics and Econometrics: Tenth World Congress, Volume III,* edited by Daron Acemoglu, Manuel Arrelano, and Eddie Deckel, 184-230. New York: Cambridge University Press.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression With Weak Instruments." *Econometrica* 65(3): 557-86.

Stock, James H., and Jonathan H. Wright. 2000. "GMM With Weak Identification." *Econometrica* 68(5): 1055-96.

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