

# Inference in models with adaptive learning

Sophocles Mavroeidis<sup>a,\*</sup> Guillaume Chevillon<sup>b,†</sup> Michael Massmann<sup>c,‡</sup>

<sup>a</sup>Brown University   <sup>b</sup>ESSEC Business School and CREST-INSEE   <sup>c</sup>Vrije Universiteit Amsterdam

September 17, 2009

---

## Abstract

We show that identification of structural parameters in models with adaptive learning can be weak, causing standard inference procedures to become unreliable. Learning also induces persistent dynamics, and this makes the distribution of estimators and test statistics non-standard. We show that valid inference can be conducted using the Anderson-Rubin statistic with appropriate choice of instruments. Application of this method to a typical new Keynesian sticky-price model with perpetual learning demonstrates its usefulness in practice.

*Keywords:* Weak identification, Persistence, Anderson-Rubin statistic, DSGE models

*JEL classification:* C1, E3

---

\*corresponding author: [sophocles\\_mavroeidis@brown.edu](mailto:sophocles_mavroeidis@brown.edu)

†This author gratefully acknowledges research support from the Economics Department, University of Oxford and a grant by CERESSEC.

‡The authors would like to thank Jörg Breitung, Norbert Christopeit, Davide Delle Monache, Stefano Eusepi, Stéphane Gregoir, Peter Howitt, Bob King, Frank Kleibergen, Albert Marcet, Adrian Pagan, Bruce Preston, Tom Sargent, Frank Schorfheide, Jim Stock, Harald Uhlig, Mark Watson, Tao Zha, and an anonymous referee, as well as the participants in the NBER summer institute and the Learning & Macroeconomic Policy Conference in Cambridge for helpful comments and discussions. We also benefited from comments received in the NBER/NSF time series conference, several Econometric Society meetings, the NESG and T2M conferences, as well as from seminar participants at Duke, Harvard-MIT, Maastricht, NYU, Nottingham, OSU, UCL and VU Amsterdam.

## 1. Introduction

This paper studies inference in structural equations involving expectations that are modeled using adaptive learning. A growing number of studies consider adaptive learning as an alternative to rational expectations, see for instance Sargent (1993), Evans and Honkapohja (2001; 2008), Orphanides and Williams (2004; 2005), Primiceri (2006), Milani (2006; 2007). Structural models with adaptive learning are self-referential and their dynamics are considerably more complicated than the dynamics under rational expectations.<sup>1</sup> As a result, little is known about the properties of structural estimation and inference in these models.

The motivation for studying this problem is twofold. On the one hand, it is well-understood that learning typically induces more persistence in the data than what is implied by models with rational expectations. In fact, one of the motivations for replacing rational expectations with adaptive learning in forward-looking models is to match the dynamics in the data without the need to introduce any intrinsic sources of persistence, which are thought of as ad hoc, see Milani (2006; 2007). On the other hand, it is well-known that forward-looking models suffer from identification problems under rational expectations, see Mavroeidis (2005), Canova and Sala (2009) and Cochrane (2007). Hence, the objective of this paper is to study the implications of those two issues, persistent dynamics and weak identification, for inference on the structural parameters of models with adaptive learning. Our main results can be summarized as follows.

First, we show that if one would like to estimate a model in which agents have bounded rationality and use a recursive learning scheme to construct expectations, one should use identification robust methods to conduct econometric inference. This is because identification pathologies that arise under rational expectations are also relevant under adaptive learning. Moreover, we show that there is one additional complication which prevents us from using standard identification robust methods for inference. Learning induces persistence in the

---

<sup>1</sup>Models with Bayesian learning, where non-fully informed agents update their beliefs about the state of the economy using Bayes rule, also induce more complex dynamics than full-information rational expectations, see Schorfheide (2005).

data and can cause nearly non-stationary behavior.

Second, we show that there is a straightforward and easy-to-implement solution to the problem of inference. In particular, we propose to use a statistic developed by Anderson and Rubin (1949) and popularized recently by the weak instruments literature, with an appropriate choice of instruments, such as lags of the identified structural shocks. The limiting distribution of this statistic is standard and does not depend on any nuisance parameters, so inference based on it is robust to identification and persistence problems. Simulations show that the method is reliable and powerful in finite samples.

Earlier studies, such as Milani (2006; 2007), adopted a Bayesian approach to inference, and primarily presented evidence on the relative fit of learning models when compared to the same models under rational expectations. From that perspective, one contribution of this paper is to provide a measure of absolute fit for models with learning using classical inference.

Finally, we apply our method to examine a new Keynesian sticky-price model with adaptive learning that has been recently studied in the learning literature. We consider specifications of the model that involve expectations over both short and long horizons, see Preston (2005b). Our results show that our proposed method is powerful enough to uncover evidence against short-horizon formulations, and produces informative inference on the parameters using a long-horizon formulation, thus demonstrating its usefulness in practice.

The rest of the paper is structured as follows. Section 2 discusses the problems of inference due to weak identification and persistence in the data. Section 3 introduces our proposed method of inference and provides simulation evidence on its size and power properties in finite samples. Section 4 provides the results of the empirical application. Technical and computational details are presented in an appendix at the end.

## 2. The problem

### 2.1. Weak identification

Suppose we wish to estimate a model with parameters  $\theta \in \Theta$ , where  $\Theta$  is the admissible parameter space. Two values of the parameters  $\theta_1, \theta_2 \in \Theta$  are observationally equivalent if the data generating process (DGP) is identical at  $\theta_1$  and  $\theta_2$ . A nonidentification region is a subset of the parameter space that contains observationally equivalent parameter values. The model is weakly identified when the true value of  $\theta$  is close to a nonidentification region, see Dufour (1997). Weak identification is known to cause standard asymptotic approximations to the distribution of estimators and test statistics to break down, inducing biases and leading to spurious inference, see Stock et al. (2002), Dufour (2003) and Andrews and Stock (2005).

Identification pathologies have been studied extensively in models with rational expectations. The early literature (Pesaran 1987) focused on characterizing the non-identification region, while more recent papers emphasized problems of weak identification, see Kleibergen and Mavroeidis (2009) and the references therein. However, to the best of our knowledge, there are no identification results in models of bounded rationality where expectations are formed using adaptive learning. This paper looks at this issue.

In a model of adaptive learning, agents are assumed to form their expectations by recursively estimating some forecasting model that represents their perceived law of motion (PLM) of the data. The resulting DGP is often called the actual law of motion (ALM). Even when the structural model is linear (or log-linearized) under rational expectations, the ALM that results with adaptive learning can be highly nonlinear. As a result, identification analysis is substantially more involved under learning than under rational expectations. Therefore, we do not attempt to provide a complete characterization of identification pathologies in models with adaptive learning. Instead, we show that identification pathologies that arise under rational expectations are relevant also in models of bounded rationality.

Let  $a_t$  denote agents' beliefs about the parameters of the PLM at time  $t$ . Under adaptive

learning,  $a_t$  evolves according to a stochastic recursive algorithm of the form:<sup>2</sup>

$$a_t = a_{t-1} + \gamma_t H(a_{t-1}, X_t) + \gamma_t^2 \rho(a_{t-1}, X_t), \quad (1)$$

where  $X_t$  is a vector of observable state variables,  $H(\cdot)$ ,  $\rho(\cdot)$  are functions describing how  $a_t$  is updated and  $\gamma_t \in (0, 1)$  is a gain parameter sequence that determines the weight agents put on new data when they update their estimates. It is often either a decreasing sequence or a constant.

The theoretical learning literature has established conditions under which rational expectations equilibria are learnable. These include restrictions on the structural parameters, known as E-stability conditions, and restrictions on the gain sequence  $\gamma_t$ . In many popular economic models, it has been shown that decreasing-gain learning converges to a REE with probability 1, and convergence results have also been established for constant gain algorithms, see Evans and Honkapohja (2001) (henceforth EH) for a review of the literature. In particular, the ALM of a model with learning can get arbitrarily close to a REE as the gain parameter gets smaller. This situation is empirically relevant, because researchers are often interested in estimating the dynamics of the economy when there are only small departures from rational expectations, see, e.g., Milani (2007).

In a model with learning, the parameter vector  $\theta$  also includes the parameters that characterize the learning algorithm. Since values of the gain parameter that are arbitrarily close to zero are within the admissible parameter space, the true value of the parameters may be arbitrarily close to a nonidentification region that arises under rational expectations. It follows that identification pathologies that exist under rational expectations will induce weak identification under adaptive learning.

When the PLM is misspecified in the sense that it does not nest any REE, learning may still converge to a so-called ‘restricted perceptions equilibrium’, see EH section 3.6.

---

<sup>2</sup>This is the most general formulation taken from equation 6.3 in Evans and Honkapohja (2001). Typically the second order term  $\rho(\cdot)$  is absent.

It is straightforward to show that identification problems can also arise under restricted perceptions equilibria. An example is given below. Therefore, we have established the following result.

**Proposition 1** *Identification of the parameters in a model with adaptive learning can be weak when learning may induce only small deviations from a rational expectations or restricted perceptions equilibrium under which the parameters of the model are not identified.*

We illustrate the above result and its implications for inference using a classic example of a model with adaptive learning.

*An example*

Consider the model

$$y_t = \beta y_t^e + \delta x_{t-1} + \eta_t \quad (2)$$

where  $\eta_t$  is an innovation process,  $y_t^e$  denotes expectations of  $y_t$  based on information available at time  $t - 1$ , and  $x_{t-1}$  is a vector of exogenous and predetermined variables. This is the model studied in the early learning literature by Bray and Savin (1986) and Fourgeaud et al. (1986). EH (section 2.2) motivate this as a reduced-form price equation arising from either a simple cobweb model or the well-known Lucas (1973) aggregate supply model.

Provided  $\beta \neq 1$ , the unique REE is given by:

$$y_t = \alpha x_{t-1} + \eta_t, \quad \alpha = (1 - \beta)^{-1} \delta. \quad (3)$$

We assume that equation (3) is the PLM, and agents' learning algorithm is either recursive least squares (RLS) or constant-gain least squares (CGLS), which are both special cases of the stochastic recursive algorithm given in equation (1) above, see EH (chapter 2). Using agents' estimate  $a_{t-1}$  of  $\alpha$  to obtain their forecast,  $y_t^e = a_{t-1} x_{t-1}$ , and substituting it into

equation (2) yields the ALM of  $y_t$ :

$$y_t = \beta a_{t-1} x_{t-1} + \delta x_{t-1} + \eta_t. \quad (4)$$

The structural parameters  $\beta$  and  $\delta$  in equation (2) are clearly not identified under rational expectations, because the regressors  $y_t^e = \alpha x_{t-1}$  and  $x_{t-1}$  are perfectly collinear. In fact, this holds for a more general class of models involving expectations of current values of the endogenous variables as regressors, see Pesaran (1987). On the other hand, learning breaks the perfect collinearity between the regressors, as long as  $a_{t-1}$  is not constant.

Turning to E-stability, it can be shown that when  $\beta < 1$  and  $x_t$  is stationary,  $a_t$  converges to  $\alpha$  under RLS learning with probability one, see EH (Theorem 2.1). Therefore, the regressors  $y_t^e = a_{t-1} x_{t-1}$  and  $x_{t-1}$  in (2) become perfectly collinear in large samples and identification breaks down. This is an example of a phenomenon known in econometrics as near multicollinearity, see, e.g., Judge et al. (1985).

Near multicollinearity will also arise when  $a_t$  converges to some fixed value other than  $\alpha$ , as in the case of a restricted perceptions equilibrium. For example, suppose  $x_t = (x_{1t}, x_{2t})$  but agents omit  $x_{2t}$  from their PLM. In the resulting restricted perceptions equilibrium  $y_t^e$  would be a linear combination of  $x_{1,t-1}$  and so it would still be perfectly collinear with  $x_{t-1}$ .

In the case of constant gain learning,  $a_t$  no longer converges to a constant, but its variability crucially depends on the gain parameter  $\gamma$ . Specifically, it can be shown that, when the E-stability conditions hold,  $a_t$  converges in probability to  $\alpha$  as  $\gamma$  tends to zero, see EH (section 14.2). Therefore, since  $\gamma$  can be arbitrarily close to zero,  $a_t$  can be arbitrary close to a constant, in which case identification will be weak.

To illustrate the severity of the problem of weak identification, we report simulation results on the finite sample properties of  $t$  and Wald tests on the parameters  $\beta$  and  $\delta$  in equation (2). We set  $x_{t-1} = 1$  and normalize the true value of  $\delta$  to zero.<sup>3</sup> We also assume the learning algorithm is known, so  $y_t^e$  is observed – the problem is even more severe when

---

<sup>3</sup>Results for multiple stochastic regressors are similar.

the parameters of the learning algorithm need to be estimated as well. Figure 1 shows the densities of the  $t$  statistics under the null for samples of size  $T = 100, 10^3$  and  $10^4$  observations, and compares these densities to the standard normal asymptotic approximation. It is clear from the pictures that the normal distribution provides a very poor approximation to the sampling distribution of the statistics even for  $T = 10^4$ . Non-normality is also evident in the distributions of the OLS estimators of  $\beta$  and  $\delta$ , which are not reported for brevity.

The pictures show that under CGLS learning there is convergence to normality. This is because when  $\gamma$  is bounded away from zero and kept fixed as the sample size gets larger, there is no multicollinearity in large samples. Therefore, the relevant question is how large the sample needs to be for the asymptotic approximations to become reliable. This question can be addressed by looking at the minimum sample size that is required for the actual rejection frequency of an asymptotic 5% level test not exceed some tolerance level, say 10%.<sup>4</sup> Table 1 reports the resulting minimum sample sizes for a  $t$  test on  $\beta$ , when  $\beta$  varies between 0.9 and 0.99 and  $\gamma$  varies between 0.1 and 0.01. The main message is that the required sample size increases as  $\beta$  gets closer to unity and  $\gamma$  closer to zero. Notably, when  $\beta = 0.99$  and  $\gamma = 0.01$ , the required sample size is around 100,000 observations!<sup>5</sup>

## 2.2. Persistence

Next, we turn to the issue of persistence induced by learning dynamics. We focus our discussion on the example introduced in the previous subsection. In equation (2) the persistence of  $y_t$  and  $y_t^e$  under the REE (3) is determined solely by the dynamics of the driving process  $x_t$ , but learning adds further dynamics to  $y_t$ . Thus, we need to examine how much persistence learning generates, and what implications this has for inference on the structural parameters.

We consider CGLS learning, since it is more relevant empirically than RLS learning. To

---

<sup>4</sup>A similar approach was used by Stock and Yogo (2003) to define weak instruments.

<sup>5</sup>Failure of conventional asymptotic theory at such large sample sizes is not unprecedented in economics. A classic example is found in Bound et al. (1995), who reported problems with the two-stage least squares estimator of the returns to education in a sample of 329,000 observations.

keep the exposition simple, we discuss only the case in which the regressor  $x_t$  is a scalar constant, because in that case, the ALM reduces to a linear time series model, which most readers are familiar with. When  $x_t = 1$  in model (2), CGLS can be expressed recursively as  $a_t = a_{t-1} + \gamma(y_t - a_{t-1})$ . Substituting for  $y_t$  using (2), the law of motion for  $a_t$  can be written as a first autoregression:

$$a_t - \alpha = (1 - (1 - \beta)\gamma)(a_{t-1} - \alpha) + \gamma\eta_t, \quad t = 1, 2, \dots \quad (5)$$

Hence, when the autoregressive root is stable, i.e., when  $|1 - (1 - \beta)\gamma| < 1$ , the process  $a_t$  admits a stationary solution and is ergodic. This implies that the asymptotic distribution theory for OLS estimators and Wald tests is standard when  $\gamma > 0$  and  $1 - 2/\gamma < \beta < 1$ .

For values of the parameters  $\gamma$  and  $\beta$  that are close to the boundaries of zero and one, respectively, it is evident from equation (5) that  $a_t$  follows an autoregressive process with a near unit root. Since it is well-known that distribution theory for nearly integrated autoregressive processes is non-standard (see Phillips, 1987), we expect that this may have an impact on the distribution of estimators and test statistics for the coefficients in equation (2), as well. To approximate the behavior of the regressor  $y_t^e = a_{t-1}$  and the OLS estimator  $(\widehat{\beta}, \widehat{\delta})' = \left(\sum_{t=1}^T X_t' X_t\right)^{-1} \sum_{t=1}^T X_t' y_t$ , where  $X_t = (y_t^e, 1)$ , we let  $\gamma$  lie in a neighborhood of zero and  $\beta$  in a neighborhood of one as the sample size grows, i.e., we set  $\gamma \in O(T^{-\nu})$  and  $1 - \beta \in O(T^{-\omega})$  with  $\nu, \omega > 0$ . This nesting is such that model parameters are constant in any given sample, but they are allowed to get closer to the boundaries as the sample grows (for notational convenience, we suppress the dependence of  $\beta$  and  $\gamma$  on  $T$  in the results given below). This approach is known as local asymptotic approximation, and it can lead to better asymptotic approximations to the finite-sample distributions of statistics that involve persistent data, see Chan and Wei (1987) and Phillips (1987).

The rates  $\nu$  and  $\omega$  characterize, respectively, the proximity to zero of  $\gamma$  and  $1 - \beta$  in terms of  $T$  and different choices for them give rise to alternative local asymptotic approximations

to the behavior of  $a_t$  and of the estimators of  $(\beta, \delta)$ . In the following proposition, we only give the results for the case  $\nu = \omega = 1/2$ , since this localization was found to give by far the most accurate asymptotic approximation to the finite sample distributions.<sup>6</sup> The symbol “ $\Rightarrow$ ” denotes weak convergence.

**Proposition 2** *Consider the stochastic process  $a_t$  that satisfies equation (5) with initial condition  $a_0$ . Suppose  $(1 - \beta)\gamma = 1 - e^{\phi/T}$  and  $\gamma = \psi/\sqrt{T}$  with  $\phi < 0$  and  $\psi > 0$ , and let  $[Tr]$  denote the integer part of  $Tr$ , for  $0 \leq r \leq 1$ . Then, as  $T \rightarrow \infty$ ,*

$$a_{[Tr]} \Rightarrow \alpha + e^{\phi r} (a_0 - \alpha) + \psi \sigma_\eta J_\phi(r) \equiv K_{\psi, \phi}(r) \quad (6)$$

where  $J_\phi(r)$  is an Ornstein-Uhlenbeck diffusion with parameter  $\phi$  and  $J_\phi(0) = 0$ , driven by the standard Brownian motion  $W(r)$ , which is such that  $T^{-1/2} \sum_{t=1}^{[Tr]} \eta_t \Rightarrow \sigma_\eta W(r)$ . Moreover, let  $(\widehat{\beta}, \widehat{\delta})' = \left( \sum_{t=1}^T X_t' X_t \right)^{-1} \sum_{t=1}^T X_t' y_t$ , where  $X_t = (y_t^e, 1)$ . Then,

$$\begin{bmatrix} \sqrt{T} (\widehat{\beta} - \beta) \\ \sqrt{T} (\widehat{\delta} - \delta) \end{bmatrix} \Rightarrow \begin{bmatrix} \int_0^1 K_{\psi, \phi}^2(r) dr & \int_0^1 K_{\psi, \phi}(r) dr \\ \int_0^1 K_{\psi, \phi}(r) dr & 1 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_\eta \int_0^1 K_{\psi, \phi}(r) dW(r) \\ \sigma_\eta W(1) \end{bmatrix}. \quad (7)$$

In the above result, the parameters  $\phi$  and  $\psi$  measure, respectively, the distance of the autoregressive root from unity and of the gain parameter from zero, relative to the sample size. The Ornstein-Uhlenbeck diffusion is a continuous time autoregressive process whose persistence is inversely related to  $\phi$ , where the limiting case  $\phi = 0$  corresponds to a random walk.s

Regarding the asymptotic distribution of the OLS estimators, we see that it is non-normal. This is because the second moment matrix of the regressors,  $T^{-1} \sum_{t=1}^T X_t' X_t$ , does not converge to a non-stochastic limit, and the sample moment conditions involving the persistent regressor,  $T^{-1/2} \sum_{t=1}^T y_t^e \eta_t$ , do not satisfy a standard central limit theorem. In the special case  $\alpha = a_0 = 0$ , the distribution of the OLS estimators given by the right-hand side

<sup>6</sup>Results for all other cases of  $\nu$  and  $\omega$  are available from the authors on request.

of equation (7) corresponds almost exactly to the local-to-unit root approximation in the model considered by Phillips (1987) and the resulting distributions are of the Dickey-Fuller type. Yet, contrary to the pure unit-root case,  $\hat{\beta}$  does not converge faster than at rate  $\sqrt{T}$ . This is because of the dampening effect of a vanishing  $\gamma$  on the variance of the regressor  $y_t^e$ , which prevents it from exhibiting a stochastic trend.

Figure 2 shows that the local asymptotic distribution given by the right-hand side of expression (7) provides a very accurate approximation to the finite sample distribution of the OLS estimators for a sample of size  $T = 100$  and for  $\beta = 0.99$  and  $\gamma = 0.02$ , while the standard fixed-parameter asymptotic approximation is poor. The approximation is also very good for other values of  $\beta$  and  $\gamma$ .

### 3. Robust inference

#### 3.1. The Anderson-Rubin statistic

In this section, we propose a method of inference that is robust to the weak identification and persistence problems we discussed above. The proposed method is an application of the Anderson and Rubin (1949) test (henceforth  $\mathcal{AR}$  test), which has been recently revived by the weak instruments literature. The original  $\mathcal{AR}$  test applies to a linear instrumental variable (IV) model with strongly exogenous instruments and Gaussian independently and identically distributed (i.i.d.) data. However, Stock and Wright (2000) extended it to nonlinear models with dependent and heterogeneous data that are estimable by the generalized method of moments (GMM), under mild regularity conditions. Here we show how to obtain versions of the  $\mathcal{AR}$  statistic for which the regularity conditions in Stock and Wright (2000) can be verified for models with learning. For a detailed description of the  $\mathcal{AR}$  statistic, the reader is referred to the excellent surveys of Stock et al. (2002), Dufour (2003) and Andrews and Stock (2005).

To explain our proposed method of inference, it is useful to illustrate the basic principle of the  $\mathcal{AR}$  test in its original setting, the prototypical linear IV model. Suppose we are

interested in testing the null hypothesis  $H_0 : \theta = \theta_0$  on the parameters of the model  $y_t = Y_t\theta + u_t$ , where the regressor  $Y_t$  is endogenous and  $u_t$  is a disturbance term, and suppose there exists an exogenous vector of instruments  $Z_t$  such that  $E(Z_t u_t) = 0$ . The principle of the  $\mathcal{AR}$  test is not to test  $H_0$  directly, but rather take a somewhat lateral approach by testing the exclusion restrictions that are implied by  $H_0$ . Since, under  $H_0$ , the disturbance term is observed,  $u_t^0 \equiv y_t - Y_t\theta_0$ , the  $\mathcal{AR}$  test can be obtained as the usual  $F$  test of  $\psi = 0$  in the auxiliary regression  $u_t^0 = \psi Z_t + \zeta_t$ . Moreover, the distribution of the  $\mathcal{AR}$  statistic does not depend on the correlation between the regressors and the instruments, i.e. on whether the parameters  $\theta$  are identified or not, and is therefore fully robust to weak instruments. When the model is just-identified, the  $\mathcal{AR}$  test even has certain optimality properties, see Andrews and Stock (2005).

The models that we consider can be expressed in the form  $h(\mathcal{Y}_t; \theta) = \eta_t$ , where  $\mathcal{Y}_t$  denotes the observed data and  $\eta_t$  is an unobserved disturbance, which could be a vector when the model consists of several equations. The parameter vector  $\theta$  includes both the structural parameters of the model and the parameters of the learning algorithm, such as the gain parameter under CGLS. Identifying assumptions, such as exclusion restrictions, are usually placed on the dynamics of the disturbance term, which is typically interpreted as a structural shock. A common assumption is that  $\eta_t$  is a martingale difference sequence, such that  $E_{t-1}\eta_t = 0$ . Based on this over-identifying assumption, we can identify the parameters by the moment conditions  $E[Z_t h(\mathcal{Y}_t; \theta)] = 0$  for any predetermined instruments  $Z_t$ . Then, the  $\mathcal{AR}$  statistic for testing the hypothesis  $H_0 : \theta = \theta_0$  is given by:<sup>7</sup>

$$\mathcal{AR}(\theta_0) = \frac{1}{T} \left( \sum_{t=1}^T \eta_t^{0'} Z_t' \right) \widehat{V}_{Z\eta}^{-1} \left( \sum_{t=1}^T Z_t \eta_t^0 \right) \quad (8)$$

where  $\eta_t^0 = h(\mathcal{Y}_t, \theta_0)$  and  $\widehat{V}_{Z\eta}$  is an estimator of the asymptotic variance of  $T^{-1/2} \sum_{t=1}^T Z_t \eta_t^0$ ,

---

<sup>7</sup>When  $\eta_t^0$  is a vector, i.e., when the model consists of a system of structural equations,  $Z_t$  must be defined as a matrix whose dimension is conformable to  $\eta_t^0$ , and the auxiliary regression is a system of seemingly unrelated regressions (SUR).

For example  $\widehat{V}_{Z\eta}$  could be White's (1980) heteroskedasticity consistent estimator, which is consistent under the assumption  $E_{t-1}\eta_t = 0$  and some additional mild regularity conditions, see Nicholls and Pagan (1983).

When the sample moments  $T^{-1/2} \sum_{t=1}^T Z_t \eta_t$  satisfy a central limit theorem, the asymptotic distribution of  $\mathcal{AR}(\theta_0)$  under  $H_0$  is  $\chi^2(k)$ , where  $k$  is the number of moment conditions. Stock and Wright (2000) provide sufficient primitive conditions to establish this result, but when  $Z_t$  is highly persistent these conditions may not hold. For example, in the model we studied in the previous section, we saw that limit theory involving the persistent regressor  $y_t^e$  is nonstandard, and this has an impact on the  $\mathcal{AR}$  statistic as well. If we were to use  $y_t^e = a_{t-1}$ , which is predetermined, as an instrument for equation (2), then we would have  $T^{-1/2} \sum_{t=1}^T Z_t \eta_t \Rightarrow \sigma_\eta \int_0^1 K_{\psi,\phi}(r) dW(r)$ , which is non-normal, see Proposition 2. Therefore, to avoid having to work out special asymptotic theory for the  $\mathcal{AR}$  statistic in every application, we need to use instruments that ensure that the  $\mathcal{AR}$  statistic is asymptotically  $\chi^2$  under  $H_0$ . This includes predetermined variables that are weakly dependent, but it excludes lags of the endogenous variable  $y_t$  or its forecast  $y_t^e$ , which depend on the recursive estimates  $a_t$ , and may therefore be highly persistent.

The key idea behind the solution we propose is the observation that the lags of  $\eta_t$  are valid instruments, and that, since  $\eta_t$  is a martingale difference sequence, the asymptotic normality of  $T^{-1/2} \sum_{t=j}^T \eta_t \eta_{t-j}$  can be established under mild conditions based on standard limit theory, see e.g., Hamilton (1994) or White (1984).<sup>8</sup> So, by suitable selection of instruments and the use of the  $\mathcal{AR}$  statistic, we have turned a difficult problem into a trivial one.

The above principle can be generalized to cover alternative assumptions on the time dependence of the disturbances,  $\eta_t$ . For example, suppose we wish to weaken the assumption  $E_{t-1}\eta_t = 0$  to the assumption that  $\eta_t$  is an AR(1) process, which is common in applied work. The  $\mathcal{AR}$  statistic (8) can be easily adapted to this alternative specification: run the auxiliary

---

<sup>8</sup>This idea is motivated by the recent work of Gorodnichenko and Ng (2007), who used a similar approach to develop methods of inference that are robust to misspecification in the way in which data has been detrended.

regression  $\eta_t^0 = \sum_{j=1}^m \psi_j \eta_{t-j}^0 + \zeta_t$  and test the hypothesis that  $\psi_j = 0$  for all  $j > 1$ .

A confidence set for the parameters  $\theta$  with confidence level  $\varphi$  can be obtained by ‘inverting’ the  $\mathcal{AR}$  test in the usual way. This set consists of all points  $\theta_0 \in \Theta$  such that the  $p$ -value associated with  $\mathcal{AR}(\theta_0)$  is greater than  $1 - \varphi$ . Confidence sets for subsets of the parameters can be obtained by projection methods. See the Appendix for computational details.

The minimum value of the  $\mathcal{AR}$  statistic,  $\min_{\theta} \mathcal{AR}(\theta)$ , serves as a formal test of the fit of the model to the data. When  $\min_{\theta} \mathcal{AR}(\theta)$  exceeds the critical value given by the  $\varphi$  quantile of the  $\chi^2(k)$  distribution, there is no value of the parameters  $\theta \in \Theta$  for which the exclusion restrictions are consistent with the data at the  $1 - \varphi$  level, and hence the  $\varphi$ -level confidence set for  $\theta$  is empty. In that case, we conclude that the model does not fit the data at the  $(1 - \varphi)$  level of significance.

Finally, the minimizer  $\hat{\theta}$  of  $\mathcal{AR}(\theta)$  can be used as a point estimate for  $\theta$ , because it has an interesting interpretation: it is the “least-objectionable” or “least-rejected” value of the parameters under the  $\mathcal{AR}$  criterion, since it is the value that results in the highest  $p$ -value associated with the  $\mathcal{AR}$  statistic. Moreover, since  $\mathcal{AR}(\theta)$  is also a continuously updated GMM objective function,  $\hat{\theta}$  can be seen as a Continuously Updated Estimator (CUE), see Stock and Wright (2000);  $\hat{\theta}$  is also a Hodges-Lehmann estimator, see Hodges and Lehmann (1963).

### 3.2. Simulations

We evaluate the finite sample size and power properties of the proposed  $\mathcal{AR}$  statistic and compare it to the Wald statistic. Our simulations are based on a forward-looking model of the form:

$$y_t = \psi_f y_{t+1}^e + \psi_b y_{t-1} + \delta x_t + \eta_t. \tag{9}$$

where  $y_{t+1}^e$  denotes expectations of  $y_{t+1}$  using information available to the agents at time  $t$ , which may include or exclude current values of the state variables.<sup>9</sup> Our empirical application

---

<sup>9</sup>See EH for a discussion of the various timing assumptions in the literature.

in the next section consists of a system of equations of this form.

The parameters of equation (9) are sometimes called ‘semi-structural’ because they can be expressed as functions of some deeper structural parameters. Here we consider the specification  $\psi_f = \beta / (1 + \beta\varrho)$  and  $\psi_b = \varrho / (1 + \beta\varrho)$ , which corresponds to the hybrid new Keynesian Phillips curve (NKPC), where  $\beta$  denotes the discount factor, and  $\varrho$  is the indexation parameter. We assume that the forcing variable  $x_t$  is an AR(1) process with innovation  $v_t$ , and the shocks  $(\eta_t, v_t)$  are drawn from a joint Normal distribution independently across time. Finally, the PLM coincides with the unique REE, and the learning algorithm is CGLS with gain parameter  $\gamma$ .

The parameters of the model can be estimated by GMM using predetermined variables as instruments. For the Wald statistic, we use the first two lags of  $y_t$  and  $x_t$  as instruments, while the  $\mathcal{AR}$  statistic is computed using two lags of  $\eta_t$  and  $x_t$  as instruments. We also allow for an unrestricted constant in the estimation and we impose the restriction that  $\beta$  is known, as is common in applied work. This is also done in order to highlight that identification problems can arise even when  $\beta$  is known. The parameter values  $\beta$ ,  $\varrho$  and  $\delta$  are chosen so as to be representative of the estimates reported in the literature, while the parameters of the forcing variable  $x_t$  are calibrated to US data.

We first study the coverage probabilities of confidence intervals derived by inverting the Wald and  $\mathcal{AR}$  tests. Table 2 displays the actual coverage probabilities for the Wald test of  $H_0 : \varrho = \varrho_0$  at nominal levels  $\varphi$  of 75%, 90%, 95% and 99%. For simplicity, we assume  $\delta$  is known. The  $\mathcal{AR}$ -based confidence sets have exact coverage, that is, a  $\varphi$ -level confidence interval for  $\varrho$  contains the true value with probability  $\varphi$ , with only slight distortions in small samples. The Wald always undercovers, which means that the usual standard error bands around the point estimate are too tight.

Figure 3 shows the power curves of the Wald and  $\mathcal{AR}$  tests of the hypothesis  $H_0 : \varrho = \varrho_0$  at the 5% nominal level of significance for sample sizes  $T = 100$  and 200. It is evident that the  $\mathcal{AR}$  test has good power, especially over the theoretically relevant parameter regions. In

particular, it rejects with high probability the null hypothesis of no indexation,  $\varrho = 0$ , and it thus can provide reliable evidence on this issue of considerable interest in applied work. The  $\mathcal{AR}$  test does not have good power for high values of  $\varrho$  against higher alternatives, e.g., it is difficult to distinguish between a high degree of indexation and complete indexation.

#### 4. Empirical application

We apply the proposed method to a new Keynesian sticky-price model of the monetary transmission mechanism, which was studied in the learning literature by Milani (2006; 2007) and Preston (2005a; 2005b; 2006), amongst others. Under rational expectations, the model's aggregate demand and supply relations can be represented by the following Euler equations:

$$\tilde{x}_t = E_t \tilde{x}_{t+1} - (1 - \beta\eta) \sigma (i_t - E_t \pi_{t+1} - r_t^n) \quad (10)$$

$$\tilde{\pi}_t = \frac{(1 - \alpha)(1 - \alpha\beta)}{\alpha(1 + \vartheta\omega)} \left[ \omega x_t + \frac{1}{(1 - \beta\eta)\sigma} \tilde{x}_t \right] + \beta E_t \tilde{\pi}_{t+1} + u_t. \quad (11)$$

where  $\tilde{\pi}_t \equiv \pi_t - \varrho\pi_{t-1}$ ,  $\tilde{x}_t \equiv (x_t - \eta x_{t-1}) - \beta\eta E_t (x_{t+1} - \eta x_t)$ ,  $\pi_t$  is the inflation rate,  $x_t$  is the output gap,  $i_t$  is the nominal interest rate, and  $r_t^n, u_t$  are exogenous shocks. The parameters  $\varrho$  and  $\eta$  measure intrinsic sources of persistence, such as indexation and habit formation, respectively;  $\sigma$  is the intertemporal elasticity of substitution;  $\alpha$  is the fraction of firms that cannot change their price in any period;  $\vartheta$  is the Dixit-Stiglitz elasticity of substitution between differentiated goods; and  $\omega$  is the elasticity of real marginal costs with respect to output.

Under RE, Equations (10)-(11) are equivalent to the optimal microfounded decision rules

given by the infinite-horizon formulation:

$$\tilde{x}_t = E_t \sum_{T=t}^{\infty} \beta^{T-t} [(1 - \beta) \tilde{x}_{T+1} - (1 - \eta\beta) \sigma (i_T - \pi_{T+1} - r_T^n)] \quad (12)$$

$$\tilde{\pi}_t = E_t \sum_{T=t}^{\infty} (\alpha\beta)^{T-t} \left[ \frac{(1 - \alpha)(1 - \alpha\beta)}{\alpha(1 + \vartheta\omega)} \left( \omega x_T + \frac{1}{(1 - \beta\eta)\sigma} \tilde{x}_T \right) + (1 - \alpha) \beta \tilde{\pi}_{T+1} + u_T \right] \quad (13)$$

but this equivalence does not hold under arbitrary subjective expectations, see Preston (2005b). We shall refer to Equations (10)-(11) as the short-horizon specification, and Equations (12)-(13) as the long-horizon specification of the model. Under adaptive learning, these two specifications have potentially different empirical implications – see Eusepi and Preston (2008) – so we study both of them.

The model is typically completed with an inertial interest rate policy rule, whose parameters are estimated jointly with the rest of the parameters of the system. Instead, we focus here only on the above two equations because of concerns that misspecification of the policy rule equation (e.g., due to regime shifts or changes in policy targets) will spill over to the estimation of the non-policy parameters.<sup>10</sup>

Our estimation results are based on quarterly US data that cover the period 1960:Q1 to 2008:Q4. Inflation is measured by the first difference in the logarithm of the implicit GDP deflator, the output gap measure is taken from the Congressional Budget Office, and interest rates are measured by the Federal Funds rate.

For the specification of agents' expectations we assume perpetual learning (CGLS) with gain parameter  $\gamma$ , with the PLM being a first-order vector autoregression (VAR) in  $\pi_t, x_t$  and  $i_t$ . This differs from Milani (2006; 2007) who includes the shocks  $u_t$  and  $r_t^n$  as additional exogenous regressors in the PLM and assumes agents have complete knowledge of their law of motion. Whether our PLM nests the REE depends on the dynamics of the exogenous

---

<sup>10</sup>In fact, we found that if we include the standard simple inertial policy rule used in Milani (2006, 2007), the model is overwhelmingly rejected.

shocks, the specification of the policy rule and the determinacy of the REE.<sup>11</sup> We abstract from these considerations here.<sup>12</sup> In terms of timing, we assume that agents do not observe  $\pi_t$ ,  $x_t$  and  $i_t$  when they form their expectations in period  $t$ , as in Milani (2006; 2007). Finally, initial beliefs are calibrated to the available pre-estimation sample, starting in 1955:Q1.<sup>13</sup>

In our estimation we fix  $\beta$  to 0.99, which is common practice, since  $\beta$  is well-identified by long-run averages of real interest rates. The parameter  $\omega$  turns out to be very poorly identified, so we fix it to the value 0.8975, as in Milani (2006), to ease the computational burden.<sup>14</sup> Similarly,  $\vartheta$ , which is not identified in the short-horizon version, is fixed to 6, which corresponds to a 20% markup of prices over marginal costs, see Sbordone (2002).

The parameter space for the estimated parameters is set, in accordance with the underlying theory, as  $0 \leq \varrho, \eta \leq 1$ ,  $0.001 \leq \alpha \leq 1$ ,  $\sigma \geq 0.001$  and the gain parameter is  $0.001 \leq \gamma \leq 0.1$ , where the upper bound is motivated by the assumptions in Milani (2006; 2007).<sup>15</sup> We use four lags of the shocks  $u_t$  and  $r_t^n$  as instruments, and also impose the restriction that  $u_t$  and  $r_t^n$  are uncorrelated, as in Milani (2006; 2007). This yields 17 identifying restrictions. We also consider a less restricted version of the model that allows  $u_t$  and  $r_t^n$  to be autoregressive of order 1, giving rise to 15 restrictions.

Estimation results are reported in Table 3. The first two columns give the results for the short-horizon specification, Equations (10) and (11). The  $\mathcal{AR}$  test indicates that this specification of the model does not fit the data at the 10% level of significance even when we allow shocks to be autocorrelated. As a result, 90% confidence intervals for the parameters

---

<sup>11</sup>For example, it can be shown that our PLM nests the REE based on the minimum state variable solution when shocks are uncorrelated and the policy rule has no more than first-order dynamics, or when shocks are autoregressive of order one and there is no intrinsic persistence, both of which are consistent with our empirical results.

<sup>12</sup>Not nesting the REE may in fact even enhance the ability of the model to fit the data, see e.g., Huang et al. (2009).

<sup>13</sup>The choice of initialization is not so important for CGLS, see Carceles-Poveda and Giannitsarou (2007), especially when the gain parameter is high, since initial beliefs are heavily discounted. In fact, our conclusions seem robust to alternative initializations.

<sup>14</sup>The poor identifiability of  $\omega$  is also evident in the results of Milani (2007), where posterior probability intervals are wider than the prior ones he used.

<sup>15</sup>Milani's prior distribution restricts the gain parameter to be less than 0.1 with probability 0.999. Most other studies fix or calibrate the gain parameters to values well below 0.1, and typically around 0.02.

are empty.

The last two columns give the results for the long-horizon specification, Equations (12) and (13). This specification fits according to the  $\mathcal{AR}$  test even without autocorrelated shocks with a  $p$ -value close to 40%. The version without autocorrelated shocks produces very tight confidence intervals for all parameters. We find that intrinsic sources of persistence are highly significant, which is in line with the corresponding results in Milani (2006, Table 1). The gain parameter  $\gamma$  is higher than is typically assumed or estimated, since values lower than 0.044 are outside the confidence interval. The price stickiness parameter  $\alpha$  is estimated quite high at 0.99 though 1 is rejected, so the Phillips curve (13) is not completely flat.<sup>16</sup> The slope of the aggregate demand relation (12), which is proportional to  $\sigma$ , is also estimated to be quite low, but again significantly different from zero. These results suggest relatively weak monetary policy transmission.

Finally, turning to the results of the model with long-horizon expectations and autocorrelated shocks, we see that the parameters are very poorly identified since all confidence intervals are uninformative. This suggests that it is difficult to distinguish intrinsic sources of persistence from exogenous dynamics in the shocks in a model with adaptive learning.

## 5. Conclusion

The objective of this paper was to study classical inference in models with adaptive learning, and we made the following contributions. First, we showed that standard methods of inference are unreliable, and we provided explanations for this result. Specifically, we showed that this can be attributed to weak identification and persistent dynamics induced by learning.

Second, we provided a solution to the problem of inference. Our proposed method is based on the Anderson and Rubin (1949) principle of testing the identifying restrictions of the model. The method is quite general and fully robust to the above problems, and its

---

<sup>16</sup>Milani (2006) reports a similar estimate, but for a model with highly autocorrelated shocks, which is the opposite of what we find here, see last column of Table 3.

implementation is straightforward. Simulations showed that our method is reliable and has good power in finite samples. An empirical application on a new Keynesian sticky-price model demonstrated the usefulness of the proposed method in practice.

## References

- Anderson, T. W. and H. Rubin (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Statistics* 20, 46–63.
- Andrews, D. W. and J. H. Stock (2005). Inference with Weak Instruments. NBER Technical Working Papers 0313, National Bureau of Economic Research, Inc.
- Bound, J., D. A. Jaeger, and R. M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association* 90(430), 443–450.
- Bray, M. M. and N. E. Savin (1986). Rational expectations equilibria, learning, and model specification. *Econometrica* 54(5), 1129–1160.
- Canova, F. and L. Sala (2009). Back to square one: Identification issues in DSGE models. *Journal of Monetary Economics* 56(4), 431 – 449.
- Carceles-Poveda, E. and C. Giannitsarou (2007). Adaptive learning in practice. *Journal of Economic Dynamics and Control* 31, 2659–2697.
- Chan, N. H. and C. Z. Wei (1987). Asymptotic inference for nearly nonstationary AR(1) processes. *Annals of Statistics* 15(3), 1050–63.
- Cochrane, J. H. (2007). Identification with Taylor Rules: A Critical Review. NBER Working Papers 13410, National Bureau of Economic Research, Inc.
- Doornik, J. A. (2007). *Object-Oriented Matrix Programming Using Ox* (Third ed.). London: Timberlake Consultants Press.

- Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica* 65(6), 1365–1387.
- Dufour, J.-M. (2003). Identification, Weak Instruments and Statistical Inference in Econometrics. *Canadian Journal of Economics* 36(4), 767–808. Presidential Address to the Canadian Economics Association.
- Eusepi, S. and B. Preston (2008). Expectations, learning and business cycle fluctuations. NBER Working Papers 14181, National Bureau of Economic Research, Inc.
- Evans, G. W. and S. Honkapohja (2001). *Learning and Expectations in Macroeconomics*. Princeton: Princeton University Press.
- Evans, G. W. and S. Honkapohja (2008). Expectations, learning and monetary policy: An overview of recent research. Discussion Paper 6640, CEPR.
- Fourgeaud, C., C. Gourieroux, and J. Pradel (1986). Learning procedures and convergence to rationality. *Econometrica* 54(4), 845–68.
- Gorodnichenko, Y. and S. Ng (2007). Estimation of DSGE models when the data are persistent. Technical report. Presented at NBER Summer Institute.
- Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.
- Hodges, J. L. and E. L. Lehmann (1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics* 34(2), 598–611.
- Huang, K., Z. Liu, and T. Zha (2009). Learning, adaptive expectations and technology shocks. *Economic Journal* 119(536), 377–405.
- Judge, G., R. Hill, W. Griffiths, H. Lutkepohl, and T.-C. Lee (1985). *The Theory and Practice of Econometrics*. New York, U.S.A.: Wiley.

- Kleibergen, F. and S. Mavroeidis (2009). Weak Instrument Robust Tests in GMM and the New Keynesian Phillips Curve. *Journal of Business and Economic Statistics* 27(3), 293–311.
- Lucas, R. E. (1973). Some international evidence on output-inflation tradeoffs. *American Economic Review* 63(3), 326–334.
- 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(3), 421–449.
- Milani, F. (2006). A Bayesian DSGE model with infinite-horizon learning: Do "mechanical" sources of persistence become superfluous? *International Journal of Central Banking* 2(3), 87–106.
- Milani, F. (2007). Expectations, learning and macroeconomic persistence. *Journal of Monetary Economics* 54(7), 2065–2082.
- Nicholls, D. F. and A. R. Pagan (1983). Heteroscedasticity in models with lagged dependent variables. *Econometrica* 51(4), 1233–42.
- Orphanides, A. and J. C. Williams (2004). Imperfect knowledge, inflation expectations, and monetary policy. In B. Bernanke and M. Woodford (Eds.), *The Inflation Targeting Debate*. University of Chicago Press.
- Orphanides, A. and J. C. Williams (2005). The decline of activist stabilization policy: Natural rate misperceptions, learning, and expectations. *Journal of Economic Dynamics and Control* 29(11), 1927–1950.
- Pesaran, M. H. (1987). *The Limits to Rational Expectations*. Oxford: Blackwell Publishers.
- Phillips, P. C. B. (1987). Towards a unified asymptotic theory for autoregression. *Biometrika* 74(3), 535–547.

- Preston, B. (2005a). Adaptive learning in infinite horizon decision problems. mimeo, Columbia University.
- Preston, B. (2005b). Learning about monetary policy rules when long-horizon expectations matter. *International Journal of Central Banking* 1(2), 81–126.
- Preston, B. (2006). Adaptive learning, forecast-based instrument rules and monetary policy. *Journal of Monetary Economics* 53(3), 507–535.
- Primiceri, G. E. (2006). Why inflation rose and fell: Policymakers' beliefs and US postwar stabilization policy. *Quarterly Journal of Economics* 121(3), 867–901.
- Sargent, T. J. (1993). *Bounded Rationality in Macroeconomics*. Oxford: Clarendon Press.
- Sbordone, A. M. (2002). Prices and unit labor costs: a new test of price stickiness. *Journal of Monetary Economics* 49, 265–292.
- Schorfheide, F. (2005). Learning and monetary policy shifts. *Review of Economic Dynamics* 8(2), 392–419.
- Stock, J. H. and J. H. Wright (2000). GMM with weak identification. *Econometrica* 68(5), 1055–1096.
- Stock, J. H., J. H. Wright, and M. Yogo (2002). GMM, weak instruments, and weak identification. *Journal of Business and Economic Statistics* 20, 518–530.
- Stock, J. H. and M. Yogo (2003). Testing for weak instruments in linear IV regression. NBER technical working paper, 284, NBER, USA.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48(4), 817–38.
- White, H. (1984). *Asymptotic Theory for econometricians*. New York: Academic Press.

Woodford, M. (2003). *Interest and Prices: Foundations of a Theory of Monetary Policy*.  
Princeton: Princeton University Press.

$\beta$	$\gamma$		
	0.01	0.05	0.1
0.90	20,000	3,000	1,000
0.95	40,000	6,000	4,000
0.99	100,000	40,000	10,000

Table 1: Estimates of the minimum sample size  $T$  that is needed for a 5% nominal level  $t$ -test on  $\beta$  to be rejected no more than 10% of the time under the null hypothesis. The model is  $y_t = \beta y_t^e + \delta + \eta_t$  with CGLS learning and gain parameter  $\gamma$ .  $\eta_t$  is Gaussian white noise with unit variance,  $\delta = 0$  and learning is initialized using a pre-sample of 1000 observations.  $T$  is incremented by  $10^n$  up to  $10^{n+1}$ , then by  $10^{n+1}$  up to  $10^{n+2}$  and so on, starting with  $n = 2$ . The number of Monte Carlo replications is 2000.

$T$	Wald				$\mathcal{AR}$			
	Confidence level $\varphi$							
	75%	90%	95%	99%	75%	90%	95%	99%
100	48.7**	63.0**	70.4**	82.0**	73.1**	88.5**	94.0**	98.6**
200	56.0**	71.2**	78.7**	89.2**	74.1*	89.4	94.4*	98.9
400	59.6**	75.5**	82.6**	92.2**	74.5	89.7	94.9	98.9
600	60.2**	76.7**	84.3**	93.1**	75.0	89.9	94.9	99.0
800	60.8**	78.3**	85.4**	94.5**	75.2	89.6	94.5*	99.0
1000	61.5**	78.2**	85.7**	94.5**	75.0	90.0	94.9	99.0
10000	66.3**	83.4**	90.4**	97.1**	75.6	90.3	95.1	99.0

Table 2: Coverage probabilities of Wald and  $\mathcal{AR}$ -based confidence sets with level  $\varphi$  for  $\varrho$  in the model  $y_t = \beta/(1 + \beta\varrho)y_{t+1}^e + \varrho/(1 + \beta\varrho)y_{t-1} + \delta x_t + \eta_t$  with CGLS learning and gain parameter  $\gamma = 0.01$ . Parameter values:  $\beta = 0.99$ ,  $\varrho = 0.65$  and  $\delta = 0.15$ .  $x_t = 0.9x_{t-1} + v_t$ , and  $(\eta_t, v_t)$  are i.i.d. Normal with zero mean and  $E(\eta_t^2) = 3$ ,  $E(\eta_t v_t) = 0.1$ ,  $E(v_t^2) = 1$ . One or two asterisks indicate coverage probability is significantly different from  $\varphi$  at the 5% or 1% level, respectively, based on 10000 Monte Carlo replications.

	Short-horizon version		Long-horizon version	
	No	Yes	No	Yes
<i>Autocorr. Shocks</i>				
Parameter				
$\rho$ (indexation)	0.39 —	0.61 —	0.85 [0.58,1]	0 [0,1]
$\eta$ (habits)	0.34 —	0 —	1 [0.93,1]	0.99 [0.03,1]
$\sigma$ (IES)	0.001 —	1.29 —	0.005 [0.0013,0.101]	0.35 [0.001,6.08]
$\alpha$ (Calvo)	1 —	0.38 —	0.99 [0.95,0.99]	0.07 [0.001,1]
$\gamma$ (gain)	0.099 —	0.099 —	0.092 [0.044,0.1]	0.064 [0.001,0.1]
$\min \mathcal{AR}(\theta)$	35.45 (0.005)	24.62 (0.056)	18.22 (0.375)	10.88 (0.761)

Table 3: Estimates of the two-equation new Keynesian sticky-price model with habits and indexation under CGLS learning. Four lags of the shocks are used as instruments in each equation, and the restriction that shocks are uncorrelated is also imposed. Models that (do not) allow for autocorrelated shocks have (17) 15 restrictions. The estimation sample is 1960:Q1-2008:Q4. Square brackets contain 90% confidence intervals based on the Anderson-Rubin statistic, parenthesis contain p-values.

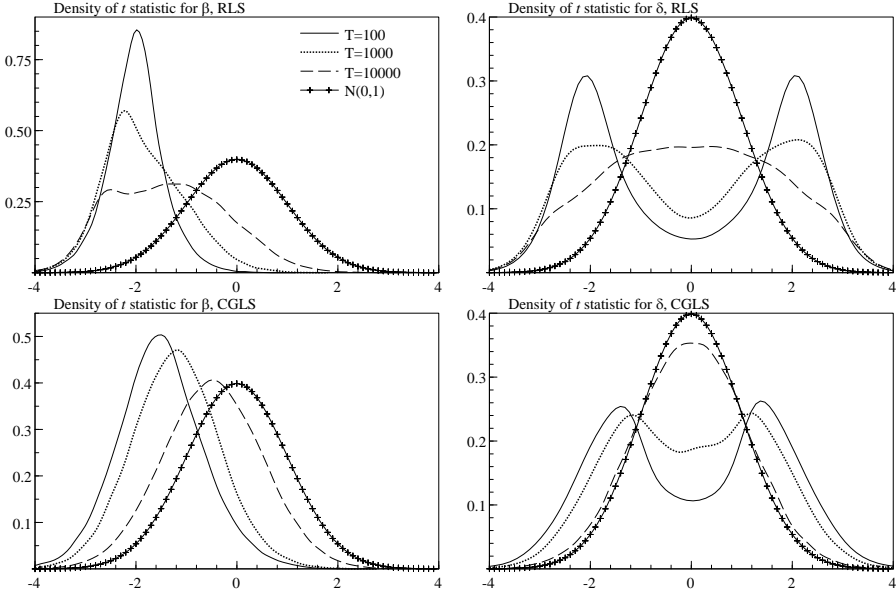


Figure 1: Densities of  $t$  statistics under the null hypothesis for the coefficients of model  $y_t = \beta y_t^e + \delta + \eta_t$ ,  $y_t^e = y_{t-1}^e + \gamma_t (y_{t-1} - y_{t-1}^e)$  for samples of size  $T = 100, 1000, 10000$ .  $\eta_t$  is Gaussian white noise with unit variance,  $\beta = 0.9$  and  $\delta = 0$ . RLS corresponds to  $\gamma_t = 1/t$ , CGLS to  $\gamma_t = 0.02$ , and  $y_0^e = 0$ . The number of MC replications is 30000.

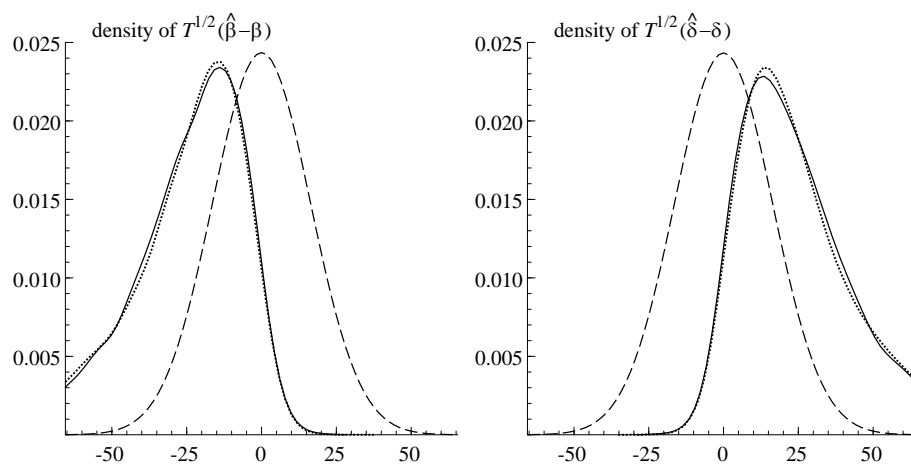


Figure 2: Densities of the OLS estimators for  $\beta$  and  $\delta$  in a sample of size  $T = 100$  (solid lines), local asymptotic approximations given by expression (7) (dotted lines) and normal asymptotic approximation (dashed lines). The model is  $y_t = \beta y_t^e + \delta + \eta_t$  with CGLS with parameter  $\gamma = 0.02$ , and  $\beta = 0.99$ ,  $\delta = 0$ , and learning is initialized at  $a_0 = 1$ . The number of MC replications is 10000.

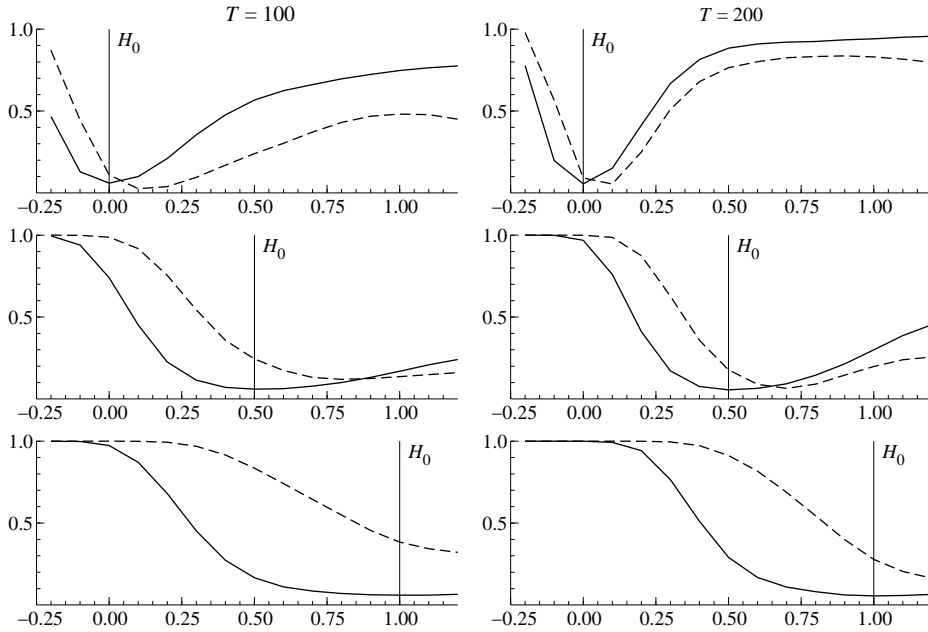


Figure 3: Power curves of 5% level Wald (dotted line) and  $\mathcal{AR}$  (solid line) tests of the null hypothesis  $H_0 : \varrho = \varrho_0$  (indicated by vertical line) in the model  $y_t = \beta/(1 + \beta\varrho)y_{t+1}^e + \varrho/(1 + \beta\varrho)y_{t-1} + \delta x_t + \eta_t$  with CGLS learning and gain parameter  $\gamma = 0.01$ . Parameter values:  $\beta = 0.99$ ,  $\varrho = 0.65$  and  $\delta = 0.15$ .  $x_t = 0.9x_{t-1} + v_t$ , and  $(\eta_t, v_t)$  are i.i.d. Normal with zero mean and  $E(\eta_t^2) = 3$ ,  $E(\eta_t v_t) = 0.1$ ,  $E(v_t^2) = 1$ . Sample size is  $T = 100$  (left column), and  $T = 200$  (right column). The number of MC replications is 10000.

## A Appendix

**Proof of proposition 2.** Solving equation (5) in terms of  $\{\eta_t\}_{t=1}^T$  and  $a_0$ , we obtain:

$$a_t - \alpha = (1 - (1 - \beta)\gamma)^t (a_0 - \alpha) + \gamma \sum_{i=0}^{t-1} (1 - (1 - \beta)\gamma)^i \eta_{t-i}.$$

Substituting for  $\beta$  and  $\gamma$  using  $\gamma = \psi/\sqrt{T}$  and  $1 - (1 - \beta)\gamma = \exp(\phi/T)$ , this can be written as:

$$a_t - \alpha = e^{\phi t/T} (a_0 - \alpha) + \frac{\psi}{\sqrt{T}} \sum_{i=0}^{t-1} e^{\phi i/T} \eta_{t-i}.$$

Let the standard Brownian motion  $W$  such that, for  $0 \leq r \leq 1$ ,  $T^{-1/2} \sum_{t=1}^{\lceil Tr \rceil} \eta_t \Rightarrow \sigma_\eta W(r)$  as  $T \rightarrow \infty$ . Then  $\frac{\psi}{\sqrt{T}} \sum_{i=0}^{\lceil Tr \rceil - 1} e^{\phi i/T} \eta_{\lceil Tr \rceil - i} \Rightarrow \psi \sigma_\eta J_\phi(r)$ , see Phillips (1987, Lemma 1), where  $J_\phi$  is an Ornstein-Uhlenbeck diffusion with  $J_\phi(0) = 0$  and parameter  $\phi$ , driven by the Brownian motion  $W(r)$ . Moreover, since  $e^{\phi r} - e^{\phi \lceil rT \rceil / T} \rightarrow 0$  as  $T \rightarrow \infty$  uniformly in  $0 \leq r \leq 1$ , equation (6) follows by Slutsky's formula for weak convergence. Now, consider the scaled OLS estimators:

$$\begin{bmatrix} \sqrt{T} (\hat{\beta} - \beta) \\ \sqrt{T} (\hat{\delta} - \delta) \end{bmatrix} = \left( \begin{bmatrix} T^{-1} \sum_{t=1}^T a_{t-1}^2 & T^{-1} \sum_{t=1}^T a_{t-1} \\ T^{-1} \sum_{t=1}^T a_{t-1} & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} T^{-1/2} \sum_{t=1}^T a_{t-1} \eta_t \\ T^{-1/2} \sum_{t=1}^T \eta_t \end{bmatrix}$$

Since  $K_{\psi,\phi}(r)$  is adapted to  $W(r)$ , it follows that  $\sum_{t=1}^T a_{t-1} \frac{\eta_t}{\sqrt{T}} \Rightarrow \sigma_\eta \int_0^1 K_{\psi,\phi}(r) dW(r)$ . Moreover, application of the continuous mapping theorem shows that  $T^{-1} \sum_{t=1}^T a_{t-1} \Rightarrow \int_0^1 K_{\psi,\phi}(r) dr$  and  $T^{-1} \sum_{t=1}^T a_{t-1}^2 \Rightarrow \int_0^1 K_{\psi,\phi}^2(r) dr$ , and hence, the result (7) follows. ■

**Computation of confidence sets** The CGLS algorithm in section 4 is given by

$$a_t = a_{t-1} + \gamma (Y_t - a_{t-1} X_t) X_t' R_t^{-1} \quad (14)$$

$$R_t = R_{t-1} + \gamma (X_t X_t' - R_{t-1}) \quad (15)$$

for  $t = 1, 2, \dots$ , where  $Y_t = (x_t, \pi_t, i_t)'$  and  $X_t = (1, x_{t-1}, \pi_{t-1}, i_{t-1})'$ . The initial conditions are obtained by least squares over the pre-estimation sample of  $t_0$  quarters,  $R_0 = \sum_{j=0}^{t_0-1} X_{-j} X_{-j}'$  and  $a_0 = \sum_{j=0}^{t_0-1} Y_{-j} X_{-j}' R_0^{-1}$ .

Projection-based confidence sets for each parameter,  $\theta^i$ , of the model are obtained by comparing the statistic  $\min_{\theta^{-i} \in \Theta^{-i}} \mathcal{AR}(\theta_0^i, \theta^{-i})$  to the appropriate critical value of the  $\chi^2(k)$  distribution for each  $\theta_0^i \in \Theta^i$ , where  $\theta^{-i}$  denotes the remaining parameters. Optimization with respect to the gain parameter is performed by grid search over range 0.001 to 0.1. Grid search is used also for  $\alpha$  in the long-horizon version of the model, over the range 0.001 to 1. Since the  $\mathcal{AR}$  criterion function is continuously differentiable with respect to the remaining parameters and the parameter space is bounded, minimization is performed using a sequential quadratic programming algorithm. The range of the remaining parameters is  $(\varrho, \eta) \in [0, 1]^2$  and  $\sigma \geq 0.001$ . All computations are carried out using Ox version 5.1, see Doornik (2007).