EFWW
7/10/08
11:00 a.m.

Inference in models with adaptive learning, with an application to the new Keynesian Phillips curve

Sophocles Mavroeidis* Brown University

Guillaume Chevillon† ESSEC, Paris

Michael Massmann‡ Vrije Universiteit Amsterdam

May 31, 2008§

Abstract

Replacing rational expectations by adaptive learning algorithms complicates the dynamics of economic models. Identification of the structural parameters is improved under learning relative to rational expectations, but it deteriorates when learning converges to rational expectations. 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 the new Keynesian Phillips curve with US data provides evidence against constant gain least squares learning.

Keywords: Weak identification, Anderson Rubin statistic, Phillips curve.

JEL: C22, E31

*corresponding author: sophocles_mavroeidis@brown.edu.

†chevillon@essec.fr.

‡mmassmann@feweb.vu.nl.

§This version is preliminary: please do not quote without the authors' consent.

1 Introduction

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

On the one hand, it is well-understood that learning typically induces more persistence in the data than what is implied by models with RE. In fact, one of the motivations for replacing RE 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 (2005, 2007). On the other hand, it is well-known that forward-looking models suffer from identification problems, see Canova and Sala (2005) Mavroeidis (2005) and Cochrane (2007a,b). 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 identification of structural models is improved under learning relative to rational expectations. The intuition for this result is simple: expectations are more variable under learning than under perfect knowledge, and this improves the accuracy of estimators in models where expectations appear as regressors. However, we also find that under decreasing or small constant gain, identification becomes weak. The problem can be expressed as near-multicollinearity in regression models, or as 'weak instruments' in models identified by exclusion restrictions. Moreover, it is shown that identification is stronger when the gain parameter is larger. Weak identification invalidates inference using conventional methods, such as Wald statistics, see Stock, Wright, and Yogo (2002). However, there is one additional complication which prevents us from using standard identification-robust methods. Learning induces persistence in the data and can cause nearly non-stationary behavior. Thus, methods that rely on normal asymptotic theory become inapplicable.

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 instru-

ments, such as lags of the identified structural shocks, so that the required regularity conditions hold. The limiting distribution of the test statistic is χ^2 and does not depend on any nuisance parameters. Simulations show that our proposed method controls size in finite samples and has reasonably good power properties, and the empirical application confirms this in practice.

Third, we apply our method to study the new Keynesian Phillips curve, a very popular model of inflation dynamics, under learning. The papers most closely related to our empirical study are those by Milani (2005, 2007). Consistently with Milani, we find that indexation is unnecessary when inflation expectations are formed by some form of adaptive learning. However, unlike Milani, we find that learning with a constant gain parameter does not fit the data, since there is evidence of shifts in the gain parameter in the US over the past fifty years. Specifically, we find that the gain parameter was significantly higher during a period of macroeconomic instability (1973 to 1987) than it was before and after that period. A learning model with an endogenously determined gain parameter may therefore be more appropriate to model the dynamics of inflation in the US.

The paper is structured as follows. Section 2 discusses the problems of inference due to weak identification and persistence in the data, with a textbook example of a model with learning from Evans and Honkapohja (2001). Section 3 introduces our proposed method and provides simulation evidence on its size and power properties in finite samples. Section 4 contains an application of the method to the new Keynesian Phillips curve with adaptive learning. Proofs and additional empirical results are given in an Appendix at the end.

The following notation is used throughout the paper: "$\xrightarrow{p}$" stands for convergence in probability, "$\Rightarrow$" for weak convergence, $a_T = O(b_T)$ means that the sequence a_T/b_T is bounded, $a_T = o(b_T)$ means $a_T/b_T \to 0$, and $O_p(\cdot)$, $o_p(\cdot)$ denote bounds in probability.

2 The problem

To fix ideas, we consider a simple model taken from Evans and Honkapohja (2001), section 14.2:

$$y_t = \beta y_t^e + \delta x_{t-1} + \eta_t \tag{1}$$

where η_t is an innovation process with variance σ_η^2, and y_t^e denotes expectations 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 by Bray and Savin (1986). Evans and Honkapohja (2001) motivate this as a reduced form price equation arising either from a simple cobweb model, or the well-known Lucas (1973)

aggregate supply model. In the former example, $\beta < 0$, while in the latter $\beta \in (0, 1)$.

Provided $\beta \neq 1$, the unique rational expectations equilibrium (REE) of the model is found to be

$$y_t = \alpha x_{t-1} + \eta_t, \quad \alpha = \frac{\delta}{1-\beta}. \tag{2}$$

Equation (2) describes the law of motion under the REE. We assume that agents perceive this as the law of motion (PLM) of y_t, but they do not know α. In order to form their forecast y_t^e, they estimate α by a_t using a stochastic recursive algorithm (SRA) with gain sequence $\{\gamma_t\}$. For instance, least squares (henceforth LS) is a SRA that can be written recursively as

$$a_t = a_{t-1} + \gamma_t \left(y_t - a_{t-1} x_{t-1}\right) x'_{t-1} R_t^{-1} \tag{3}$$

$$R_t = R_{t-1} + \gamma_t \left(x_{t-1} x'_{t-1} - R_{t-1}\right) \tag{4}$$

for $t = 1, 2, ...$, given some initial conditions a_0, R_0. Two well-studied versions of LS learning are recursive least squares (RLS), obtained from (3) and (4) with $\gamma_t = 1/t$, and constant gain (CGLS) with $\gamma_t = \gamma \in (0, 1)$. The latter is also sometimes referred to as perpetual learning (see, e.g., Orphanides and Williams (2005a)) and is particularly popular in empirical work. Agents' forecasts are then given by

$$y_t^e = a_{t-1} x_{t-1}. \tag{5}$$

Equation (2) is the PLM.[1] The dynamics of y_t under learning are characterized by the so-called Actual Law of Motion (ALM)

$$y_t = \beta a_{t-1} x_{t-1} + \delta x_{t-1} + \eta_t \tag{6}$$

which is derived by substituting $a_{t-1} x_{t-1}$ for y_t^e in the structural model (1). It is clear that the dynamics of y_t under the ALM (6) are more complicated than under the REE (2). Here, we are interested in the implications of the learning dynamics for inference on the structural parameters β and δ.

[1] The assumption that the PLM coincides with the REE is inessential for the ensuing results, since they apply also under mis-specified learning, as defined in EH section 3.6.

We consider the ordinary least squares (OLS) estimator of (β, δ):[2]

$$\begin{pmatrix} \widehat{\beta} - \beta \\ \widehat{\delta} - \delta \end{pmatrix} = \underbrace{\left[\begin{pmatrix} \sum_{t=1}^{T} a_{t-1}^2 x_{t-1}^2 & \sum_{t=1}^{T} a_{t-1} x_{t-1}^2 \\ \sum_{t=1}^{T} a_{t-1} x_{t-1}^2 & \sum_{t=1}^{T} x_{t-1}^2 \end{pmatrix} \right]^{-1}}_{A_T} \underbrace{\begin{pmatrix} \sum_{t=1}^{T} a_{t-1} x_{t-1} \eta_t \\ \sum_{t=1}^{T} x_{t-1} \eta_t \end{pmatrix}}_{b_T}. \tag{7}$$

Consistency and asymptotic normality of $\widehat{\beta}, \widehat{\delta}$ require that the matrix A_T, scaled appropriately, should be invertible with probability (approaching) one. This is the rank condition for the identification of β, δ. To establish asymptotic normality at rate $\sqrt{T}$, we need conditions that guarantee $T^{-1}A_T$ converges in probability to a nonstochastic and invertible matrix and that the process $T^{-1/2}b_T$ in equation (7) satisfies a central limit theorem. Under these conditions, the OLS estimator $\widehat{\beta}, \widehat{\delta}$ and the associated t statistics are asymptotically normal, and the Wald statistics are asymptotically χ^2, under the null hypothesis. So, our question of interest is whether these asymptotic results hold and whether they provide a good approximation to the distributions of the statistics in finite samples.

We start by reporting some Monte Carlo simulations on the distribution of OLS estimators and test statistics for the model (1). For simplicity, we make the regressor x_{t-1} in the model a scalar constant, i.e., $x_{t-1} = 1$, and we normalize the true value of the coefficient δ to zero. Figures 1 and 2 show, for samples of size 100, 1000 and 10000 observations, the densities of the OLS estimators $\widehat{\beta}, \widehat{\delta}$, and compare those densities to normal approximations. It is clear from those pictures that the normal distribution provides a very poor approximation to the sampling distribution of the OLS estimators even for samples of 10000 observations. Similar results can be obtained for the distribution of the t statistics for β and δ under the null hypothesis. Their distributions are non-normal, and, in the case of the t statistic for δ, even bimodal. The graphs are omitted for brevity.

The above results suggest that there appears to be some convergence to normality under CGLS. So, a relevant question is to look at how large the sample needs to be for the asymptotic approximations to become accurate. We answer this question by looking at the distance of the distribution of the Wald statistic on β (the square of the t statistic) under the null, from its asymptotic distribution which is $\chi^2(1)$. We shall use the Kolmogorov-Smirnov statistic for equality of two distributions as a measure, and formal test, of the quality of the asymptotic approximation.[3] Table 1 reports the

[2] This is also the maximum likelihood estimator under Gaussian and homoskedastic innovations η_t.

[3] The Kolmogorov-Smirnov statistic is equal to the maximum absolute difference between two distribution functions F_1, F_2, over all the points of support in the sample, scaled by the square root of the sample size. Its use in measuring

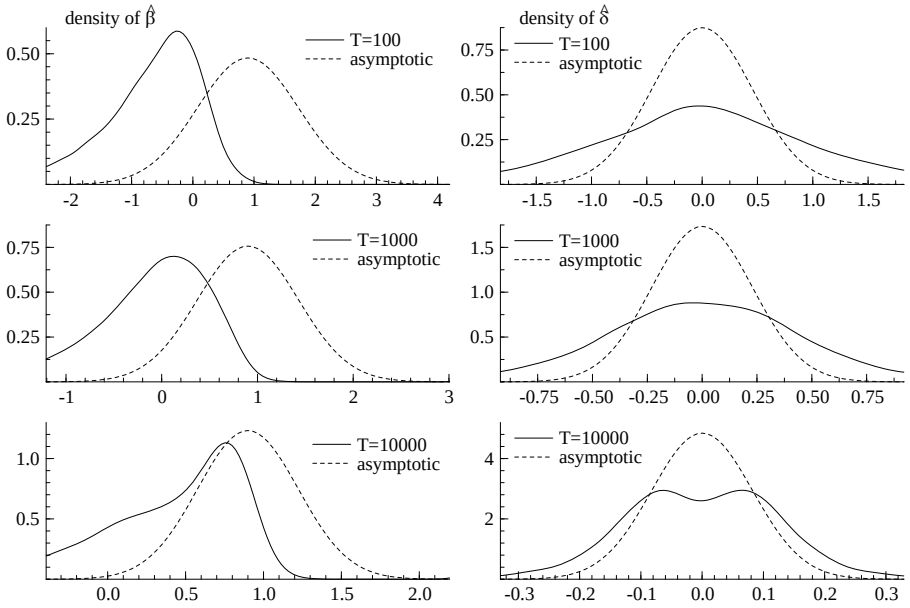

Figure 1: Densities of OLS estimators of the coefficients of model $y_t = \beta y_t^e + \delta + \eta_t$, under recursive least squares learning, for samples of size $T = 100, 1000, 10000$. η_t is Gaussian white noise with unit variance, $\beta = 0.9$ and $\delta = 0$. The number of MC replications is 10000.

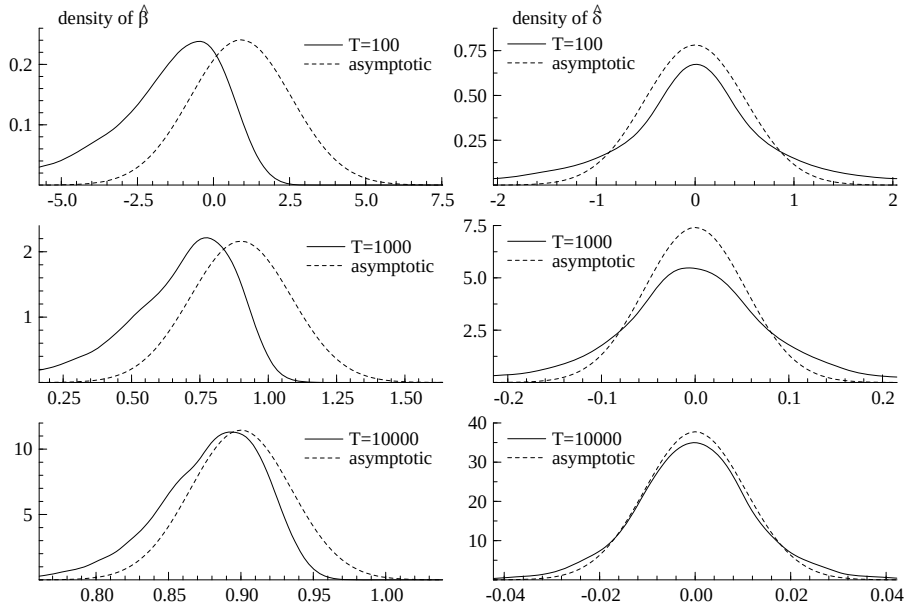

Figure 2: Densities of OLS estimators of the coefficients of model $y_t = \beta y_t^e + \delta + \eta_t$, under constant gain least squares learning, for samples of size $T = 100, 1000, 10000$. η_t is Gaussian white noise with unit variance, $\beta = 0.9$, $\delta = 0$ and $\gamma = 0.02$. The number of MC replications is 10000.

$\gamma =$	0.01	0.05	0.1
β			
0.90	100×10^3	30×10^3	5×10^3
0.95	170×10^3	50×10^3	20×10^3
0.99	500×10^3	170×10^3	80×10^3

Table 1: The table reports the minimum sample size T that is needed for the distribution of the Wald statistic on β not to be significantly different from $\chi^2(1)$ at the 5% level according to the Kolmogorov-Smirnov test. The model is $y_t = \beta y_t^e + \delta + \eta_t$, under constant squares learning with gain parameter γ. η_t is Gaussian white noise with unit variance, $\delta = 0$ and learning is initialized at zero. T is incremented by 100 up to 10000, and by 10000 thereafter. The number of Monte Carlo replications is 10000.

smallest sample sizes that are needed for the distribution of the Wald statistic to be approximately $\chi^2(1)$, for different values of β and the gain parameter γ. It is noteworthy that when $\beta = 0.99$ and $\gamma = 0.01$, the required sample size is half a million observations!

We now show that these non-standard distributions are the result of identification problems and persistence in the data, taking each explanation in turn.

2.1 Identification

It is immediately obvious from equation (2) that the parameters β and δ are not separately identified under rational expectations. One way to see this identification problem is to observe that, under the REE, the regressor y_t^e is perfectly collinear with the regressors x_{t-1}. In contrast, under learning, $y_t^e = a_{t-1}x_{t-1}$, and this breaks the perfect collinearity with x_{t-1} as long as a_{t-1} varies with t. So, learning *improves* the identifiability of the structural parameters relative to the REE.

The above discussion shows that the identification of the structural parameters β and δ hinges upon the behavior of a_t. The latter is a well-studied problem in the learning literature. For the simple model (1) with $x_t = 1$, it can be shown that provided $\beta < 1$, agents' estimator a_t converges to α under RLS learning with probability one, see Evans and Honkapohja (2001, Theorem 2.1). Hence, the regressors $y_t^e = a_{t-1}x_{t-1}$ and x_{t-1} in (1) become perfectly collinear in large samples,[4] and this is an example of a phenomenon known in econometrics as near multicollinearity, see, e.g., Judge *et al* (1985). In other words, under RLS the identification of the coefficients β and δ breaks down, and this explains the lack of convergence of the OLS estimators shown in Figure 1.

the quality of asymptotic approximations is common in econometrics, see, e.g., Staiger and Stock (1997).

[4] This is also true if a_t converges to some value other than α, as in the case of self-confirming equilibria under mis-specified learning.

The conditions under which RLS learning converges to the REE (or to some other self-confirming equilibrium under misspecification of the PLM) are referred to as E-stability conditions. These are restrictions on the structural parameters, e.g., $\beta < 1$ in the cobweb model (1), and regularity assumptions on the process x_t, see, e.g., Fourgeaud, Gourieroux, and Pradel (1986). Thus, we see that when the structural parameters of the model are not identified under the REE, RLS will lead to weak identification when the E-stability conditions hold.

With constant gain learning, it is well-known that a_t does not converge to a nonstochastic limit. Evans and Honkapohja (2001, chapter 7) discuss the behavior of a_t under constant gain learning for a large class of SRAs that includes CGLS as a special case. They show that when the constant gain parameter γ is small and $\beta < 1$, $a_t - \alpha = O_p\left(\gamma^{1/2}\right)$. Hence, it is clear that if we let γ tend to zero, there will be near multicollinearity in equation (1). This situation is in fact empirically relevant, because researchers are often interested in estimating the dynamics of the economy when there are only small departures from rational expectations, i.e., when γ is small (e.g., Milani 2007).

When γ is bounded away from zero, the multicollinearity problem disappears. This explains why in Figure 2 there appeared to be convergence under CGLS, since γ was kept fixed as we increased T. In fact, since the variability of a_t is increasing in γ, and since the accuracy with which the coefficients β, δ in (1) can be estimated is positively related to the variability of the regressors, other things equal, the parameters will be better identified (i.e., more accurately estimable) the higher is γ. In other words, under constant gain learning, identification improves as the speed of learning decreases. An illustration of this point is provided by simulations reported in section 3 below in the context of a forward-looking model, where it is shown that, for inference on the structural parameters, the gain parameter plays a role similar to the sample size.

In models that are identified by exclusion restrictions, and typically estimated by instrumental variables, decreasing or small constant gain learning leads to the problem of 'weak instruments', as it was defined by Staiger and Stock (1997). To see this, consider a model with non-predetermined regressors x_t:

$$y_t = \beta y_t^e + \delta x_t + \eta_t \tag{8}$$

Under the assumption that $E_{t-1}\eta_t = 0$, the parameters (β, δ) in equation (8) can be estimated by instrumental variables regression, using any variables known at data $t-1$ as instruments. Now, observe that the REE of (8) is given by

$$y_t = \alpha E_{t-1} x_t + \eta_t, \quad \alpha = (1-\beta)^{-1}\delta \tag{9}$$

so it is clear that (β, δ) are not identified under RE. Let z_t denote the set of instruments for predicting x_t and suppose that $E_{t-1}x_t = \kappa z_t$. Assume agents' forecasts are given by $y_t^e = \phi_t z_t$, where ϕ_t is a recursive estimate of ϕ. By equation (9), the RE forecast is $\alpha\kappa z_t$, and, if ϕ_t converges to $\alpha\kappa$, then $y_t^e = \alpha\kappa z_t + O_p(\gamma_t)$, i.e., it is asymptotically collinear with the projection of x_t on the instruments z_t. Another way to put this is that the covariance matrix between the regressors (y_t^e, x_t) and the instruments z_t becomes rank deficient as the gain parameter goes to zero.

2.2 Persistence

Next, we turn to the issue of persistence of the data under learning dynamics. We first observe that in the simple model (1) the persistence of y_t and y_t^e under the REE (2) is determined solely by the dynamics of the driving process x_t, but learning adds further dynamics to y_t independently of x_t. Thus, we need to examine how much persistence learning generates, and what implications this has for inference on the structural parameters.

We shall focus our discussion on CGLS learning, since it is more relevant empirically than RLS learning. To 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. Even though our analysis can be generalized to allow for stochastic and multiple regressors, such extensions do not add any new insights to our understanding of the problem. Moreover, the asymptotic approximations we derive below are only used to explain why standard asymptotic theory fails, as we saw in figure 2 above, and they are not used to propose solutions to the problem of inference. The solution we propose in the next section is, in fact, quite general, and does not rely on any non-standard asymptotic theory.

When $x_t = 1$ in model (1), it follows that $R_t = 1$ for all t in (4), and the SRA reduces to $a_t = a_{t-1} + \gamma(y_t - a_{t-1})$. Substituting for y_t using (1) and the fact that $y_t^e = a_{t-1}$, the law of motion for a_t can be written as a first-order autoregression with autoregressive coefficient $1-(1-\beta)\gamma$ and scale parameter γ :

$$a_t - \alpha = (1 - (1-\beta)\gamma)(a_{t-1} - \alpha) + \gamma\eta_t, \quad t = 1, 2, ... \tag{10}$$

Hence, when $\beta < 1$ and $\gamma > 0$, the process a_t is ergodic and admits a stationary solution, and this implies that the asymptotic distribution theory for OLS estimators and Wald tests is standard.

Now, let us consider what happens when the gain parameter, γ, is small. To approximate the

distribution of the stochastic process a_t and the OLS estimators in (7), we let γ lie in a neighborhood of zero, i.e., we set $\gamma = O\left(T^{-\nu}\right)$, with $\nu > 0$. We also let $\beta = 1 - O\left(T^{-\omega}\right)$ in order to characterize the situation in which β is close to one, which is often empirically relevant, e.g., when β is a discount factor. This approach leads to local asymptotic approximations, which have been used effectively to characterize the behavior of nearly integrated autoregressive processes. Note that the standard autoregression with a near unit root, see Chan and Wei (1987) and Phillips (1987), is a special case of the model (10) with $1 - \beta = O\left(T^{-1}\right)$ and γ fixed. However, since γ also affects the variance of the innovation to a_t through the term $\gamma\eta_t$ in (10), when $\gamma \to 0$, the present problem is different from the nearly integrated autoregressive model studied in the literature. e.g., in Phillips (1987), because a_t here is $O_p(1)$ rather than $O_p\left(\sqrt{T}\right)$. This has implications for the rate of convergence of the OLS estimators of $\widehat{\beta}$ and $\widehat{\delta}$ that we discuss below.

Different choices of the rates ν and ω at which γ and $1-\beta$ go to zero with T, respectively, give rise to alternative local asymptotic approximations to the behavior of a_t and of the OLS estimators $\widehat{\beta}, \widehat{\delta}$. We shall discuss here only the case $\nu = \omega = 1/2$, since this localization was found to give the best approximation to the finite sample distributions..The results are given in the following proposition.

Proposition 1 *Consider the stochastic process a_t that satisfies equation (10) 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 $T \to \infty$*

$$a_{[Tr]} \Rightarrow \alpha + e^{\phi r}(a_0 - \alpha) + \psi\sigma_\eta J_\phi(r) \stackrel{def}{=} K_{\psi,\phi}(r) \tag{11}$$

where $J_\phi(r)$ is an Ornstein-Uhlenbeck diffusion with parameter ϕ and $J_\phi(0) = 0$, driven by the standard Brownian motion $W(r)$. Moreover, the asymptotic distribution of the OLS estimators $\widehat{\beta}$, $\widehat{\delta}$ defined in equation (7) with $x_t = 1$ is

$$\begin{bmatrix} \sqrt{T}\left(\widehat{\beta} - \beta\right) \\ \sqrt{T}\left(\widehat{\delta} - \delta\right) \end{bmatrix} \Rightarrow \begin{bmatrix} \int_0^1 K_{\psi,\phi}^2(r)\,dr & \int_0^1 K_{\psi,\phi}(r)\,dr \\ \int_0^1 K_{\psi,\phi}(r)\,dr & 1 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_\eta \int_0^1 K_{\psi,\phi}(r)\,dW(r) \\ \sigma_\eta W(1) \end{bmatrix}. \tag{12}$$

In the above result, the parameters ϕ and ψ 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 ϕ, where the limiting case $\phi = 0$ corresponds to a random walk. Proposition 1

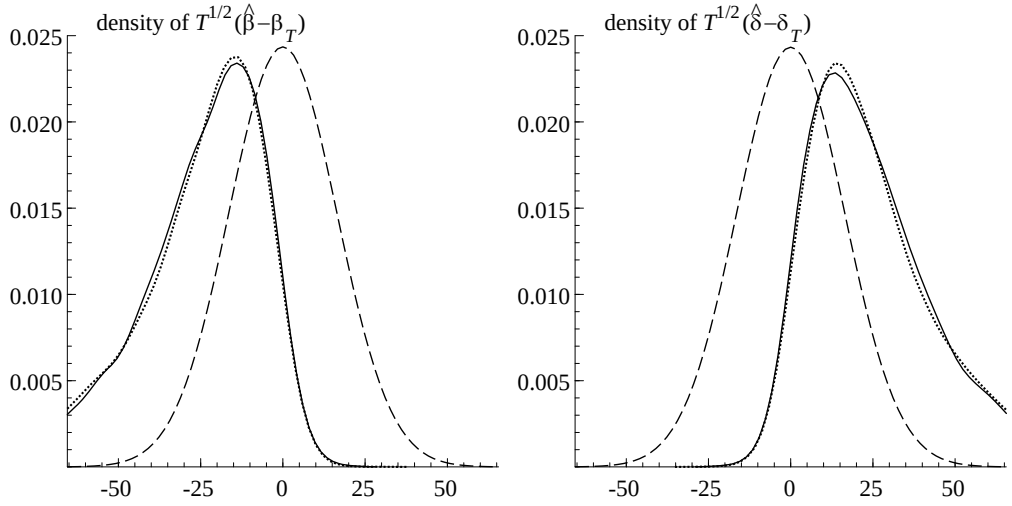

Figure 3: Densities of the OLS estimators for β and δ in a sample of size $T = 100$ (solid lines), local asymptotic approximations given by expression (**??**) (dotted lines) and normal asymptotic approximation (dashed lines). The model is $y_t = \beta y_t^e + \delta + \eta_t$ with CGLS with parameter $\gamma = 0.02$, and $\beta = 0.99$, $\delta = 0$, and learning is initialized at $a_0 = 1$. The number of MC replications is 10000.

therefore shows that the persistence in a_t is increasing the closer is $(1 - \beta)\gamma$ to zero. We also notice that a_t is not ergodic, since it does not converge to its stationary distribution for arbitrary initial condition a_0.

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 does not converge to a non-stochastic limit, and the moment conditions involving the persistent regressor a_{t-1} do not satisfy a normal central limit theorem. In the special case $\alpha = a_0 = 0$, the distribution of the OLS estimator given by the right-hand side of equation (12) corresponds almost exactly to the local-to-unit root approximation in the model considered by Phillips (1987) and the resulting distribution is of the Dickey-Fuller type, i.e. skewed towards negative values. Yet, contrary to the pure unit-root case, $\widehat{\beta}$ does not converge faster than at rate $\sqrt{T}$. This is because of the dampening effect of a vanishing γ on the variance of the regressor a_{t-1}, which we mentioned earlier.

Figure 3 shows that the local asymptotic distribution given by the right-hand side of expression (12) provides a very accurate approximation to the finite sample distribution of the OLS estimators (7) for a sample of size $T = 100$ and for $\beta = 0.99$ and $\gamma = 0.02$, for which the standard fixed-parameter asymptotic approximation is poor. The approximation is also very good for other values of β and γ, the results being omitted for brevity.

3 Robust inference using the Anderson-Rubin statistic

The previous section showed that weak identification and persistence of the regressors in models with learning render inference using conventional test statistics, such as the Wald statistic, unreliable. In this section, we propose a test statistic whose asymptotic distribution under the null is χ^2-distributed without any assumptions on identification or weak dependence in y_t. Hence, inference based on this statistic is fully robust to violations or near violations of these conditions. The proposed method is an application of the Anderson and Rubin (1949) statistic, which has been recently revived by the weak instruments literature, see Dufour (1997), Staiger and Stock (1997). The exact Anderson-Rubin (AR) statistic applies to a linear instrumental variable model with strongly exogenous instruments and Gaussian independently and identically distributed (i.i.d.) data, but 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 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 Anderson-Rubin statistic, the reader is referred to the excellent surveys of Stock et al. (2002), Dufour (2003) and Andrews and Stock (2005).

The main drawback of the AR test is that it is less powerful than the Wald test of H_0 when the regularity conditions for the latter hold, so the AR test trades off power for robustness, see Andrews and Stock (2005). Moreover, in linear models with i.i.d. data and a single endogenous regressor, Andrews, Moreira, and Stock (2006) show that another identification-robust statistic, known as the conditional likelihood ratio (CLR) statistic proposed by Moreira (2003), dominates the AR statistic in terms of power. Unfortunately, neither this, nor the score statistic proposed by Kleibergen (2005) can be used when the regressors or the instruments are highly persistent, because the conditions under which their asymptotic distribution was derived, see (Kleibergen, 2005, Assumption 1), cannot be verified. In fact, for the model we examined in the previous section, we saw that those conditions do not hold. However, we can still make valid inference using the AR statistic because the conditions for its validity are milder than those for the CLR and the score statistic.

Consider a generic model defined by the equation $h(Y_t; \theta_0) = \eta_t$, where Y_t denotes the data, θ is a vector of parameters, and η_t is an unobserved process, which could be a vector, e.g., in a multiple-equation model. Identifying assumptions are usually placed on the dynamics of the disturbance

term, e.g., $E_{t-1}\eta_t = 0$ whenever η_t is a shock. The model discussed in the previous section, see equation (1), fits in this framework, as do many popular dynamic stochastic general equilibrium models. Using the over-identifying assumption $E_{t-1}\eta_t = 0$, we can identify the parameters by the moment conditions $EZ_t h(Y_t;\theta)$ for any vector of predetermined instruments Z_t.

Consider now the problem of testing the hypothesis $H_0 : \theta = \theta_0$, against $H_1 : \theta \neq \theta_0$. Under the null hypothesis, the disturbances η_t of the model are identified by the function $h(Y_t;\theta_0)$. Hence, H_0 implies $H_0^* : EZ_t\eta_t = 0$. The AR statistic for H_0 is then the Wald statistic for testing the hypothesis H_0^*, which can be computed by running a regression of η_t (which is observable under H_0) on Z_t and testing that the coefficients of Z_t are all zero, i.e.

$$AR(\theta_0) = \frac{1}{T}\left(\sum_{t=1}^{T}\eta_t Z_t'\right)\widehat{V}_{Z\eta}^{-1}\left(\sum_{t=1}^{T}Z_t\eta_t\right) \tag{13}$$

where $\widehat{V}_{Z\eta}$ is an estimator of the variance of $T^{-1/2}\sum_{t=1}^{T}Z_t\eta_t$, such as 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).

Now, under the high level assumption that $T^{-1/2}\sum_{t=1}^{T}Z_t\eta_t$ is asymptotically normal with zero mean, the distribution of the AR statistic is, in large samples, $\chi^2(k)$ under H_0, where k is the dimension of the instrument vector Z_t. Stock and Wright (2000) discuss sufficient primitive conditions to establish this result, but when Z_t is highly persistent these conditions may not hold. As we saw in the previous section, limit theory involving the persistent regressor is nonstandard, and this has an impact on the AR statistic as well.[5] Therefore, to avoid having to work out special asymptotic theory for the AR statistic, we need to use in Z_t processes for which the asymptotic normality assumption for $T^{-1/2}\sum_{t=1}^{T}Z_t\eta_t$ can be verified. 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. In fact, a set of valid instruments is the lags of the disturbance η_t. This is motivated by the recent work of Gorodnichenko and Ng (2007), who suggested a similar approach for developing tests that are robust to misspecification in the detrending method used. Since η_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). So, by suitable selection

[5] In the model we studied in section 2, we found that $T^{-1/2}\sum_{t=1}^{T}a_{t-1}\eta_t \Rightarrow \sigma_\eta\int_0^1 K_{\psi,\phi}(r)\,dW(r)$, which is non-normal, see Proposition 1 and the Appendix.

of instruments and the use of the 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. In particular, suppose that the shock η_t is assumed to be autocorrelated. Clearly, some structure must be placed on the autocorrelation of η_t for the model to be identified, as is true for any dynamic model that does not have only strictly exogenous regressors.[6] So, suppose one makes the (relatively common) assumption that shocks are autoregressive of order one, i.e., $\eta_t = \rho_\eta \eta_{t-1} + \varepsilon_t$, where ε_t is now the underlying martingale difference process. The AR statistic (13) can be easily adapted to deal with this alternative specification. Simply run the regression of η_t on $\eta_{t-1}, ..., \eta_{t-m}$ and any n additional instruments $z_{2,t}$, and compute the Wald test for the hypothesis that all coefficients, except that on η_{t-1} are equal to zero. Under mild assumptions,[7] the asymptotic distribution of this statistic will be $\chi^2(m+n-1)$. That reasoning can, of course, be extended to η_t following any other autoregressive moving average (ARMA) process, as one sometimes encounters in applied work, see e.g., Smets and Wouters (2007).

3.1 Simulations

We evaluate the finite sample size and power properties of the proposed AR statistic and compare it to the Wald statistic, using simulations of the hybrid NKPC model of inflation. i.e. the model that we use in our empirical application in Section 4. The present section provides a brief summary of our simulation results which illustrate that our proposed method works well. Extensive simulation results are available on request.

The hybrid NKPC model with indexation takes the form

$$\pi_t = \frac{\beta}{1+\beta\varrho}\pi^e_{t+1} + \frac{\varrho}{1+\beta\varrho}\pi_{t-1} + \frac{\lambda}{1+\beta\varrho}x_t + \frac{\lambda}{1+\beta\varrho}\varepsilon_t, \tag{14}$$

where π_t denotes inflation, x_t is an observable forcing variable, and ε_t is a disturbance term. Details of the model are given in the next section. The observable forcing variable x_t is assumed to follow a second-order autoregressive process $x_t = \rho_1 x_{t-1} + \rho_2 x_{t-2} + v_t$, where the shocks ε_t, v_t are independently and jointly normally distributed with zero mean, and variance matrix $E\varepsilon_t^2 = \sigma_\varepsilon^2$,

[6] It is particularly true for models with expectations, such as dynamic stochastic general equilibrium models, see e.g., the examples discussed in Cochrane (2007a) and Beyer and Farmer (2007).

[7] For example, $\varepsilon_t, z_{2,t}$ are stationary and ergodic with finite fourth moments and $|\rho_\eta| < 1$.

$E\varepsilon_t v_t = \sigma_{\varepsilon v}$ and $Ev_t^2 = 1$. The rational expectation of π_{t+1} is given by

$$E\left(\pi_{t+1}|\pi_{t-1}, ..., x_t, x_{t-1}, ...\right) = \alpha_1\pi_{t-1} + \alpha_2 x_t + \alpha_3 x_{t-1} \tag{15}$$

where the parameters $\alpha = (\alpha_1, \alpha_2, \alpha_3)'$ are functions of the structural parameters, see Mavroeidis (2005). We assume the PLM is given by (15), and α is estimated by CGLS, so that π_{t+1}^e is given by:[8]

$$\pi_{t+1}^e = a_{1,t-1}\pi_{t-1} + a_{2,t-1}x_t + a_{3,t-1}x_{t-1} = a_{t-1}z_t.$$

Our information assumptions and the fact that $E_{t-1}\eta_t = 0$ imply that the parameters of equation (14) can be estimated by two stage least squares (2SLS) using predetermined variables as instruments. For the Wald statistic, we use the first two lags of π_t and x_t as instruments, while the AR statistic is computed using two lags of η_t and x_t as instruments. We also allow for an unrestricted constant in the estimation and we impose the restriction that β is known, as is common in applied work. With this restriction, the NKPC can be written in the following linear form:

$$y_t = \varrho w_t + \lambda x_t + \varepsilon_t \tag{16}$$

where $y_t = \pi_t - \beta\pi_{t+1}^e$, $w_t = \pi_{t-1} - \beta\pi_t$ are both endogenous regressors. The parameter values in the DGP are chosen so as to be representative of the estimates reported in the literature, e.g., Galí and Gertler (1999), while the parameters of the forcing variable x_t are calibrated to US data, see Mavroeidis (2005) for details.

We first compare the finite-sample distributions of the AR and Wald statistics for a joint test on ϱ and λ under the null with their $\chi^2(4)$ and $\chi^2(2)$ asymptotic counterparts, respectively, using the Kolmogorov-Smirnov statistic. Table 2 reports the results for different sample sizes T. The distribution of the AR statistic does not differ significantly from a $\chi^2(4)$ for most sample sizes, while the opposite is true for the Wald statistic.

Next, we study the coverage probabilities of confidence intervals derived by inverting the Wald and AR statistics. Table 3 displays the actual coverage probabilities for the Wald test of $H_0 : \varrho = \varrho_0$ at nominal levels of $\alpha = 75\%$, 90%, 95% and 99%. For simplicity, we assume λ is known. The AR-based confidence sets have exact coverage, with only slight distortions in small samples. The Wald

[8] In excluding π_t from the information set used to forecast π_{t+1} we follow the vast majority of the literature, in order to avoid the simultaneity induced by having π_t on both sides of the model (14). This informational assumption is used to simplify the simulations, and it actually makes no difference to the empirical results on the NKPC reported later.

T	Wald	AR
100	0.2266**	0.0194**
200	0.1325**	0.0101
400	0.0906**	0.0093
600	0.0712**	0.0076
800	0.0631**	0.0067
1000	0.0554**	0.0038
10000	0.0148*	0.0070

Table 2: Kolmogorov-Smirnov (KS) tests of equality of the distribution of the AR and Wald statistics to $\chi^2(4)$ and $\chi^2(2)$, respectively, under the joint null hypothesis that both ϱ and λ are equal to their true values. The parameters in the DGP are $\beta = 0.99$, $\gamma = 0.01$, $\varrho = 0.65$, $\lambda = 0.15$, $\sigma_\varepsilon = 3$, $\sigma_{\varepsilon v} = 0.1$, $\rho_1 = 0.9$ and $\rho_2 = 0$. The number of Monte Carlo replications is $M = 10000$. One asterisk denotes significance of the KS test at the 5% level, two asterisks indicate significance at the 1% level, the critical values being 0.0136 and 0.0163, respectively.

	Wald				AR			
T	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 3: Coverage probabilities of the Wald and the AR-based confidence sets with confidence levels 75%, 90%, 95% and 99% for the null hypothesis that ϱ is equal to its true value. The parameter values in the DGP are $\beta = 0.99$, $\gamma = 0.01$, $\varrho = 0.65$, $\lambda = 0.15$, $\sigma_\varepsilon = 3$, $\sigma_{\varepsilon v} = 0.1$, $\rho_1 = 0.9$ and $\rho_2 = 0$. The number of Monte Carlo replications is $M = 10000$. One or two asterisks indicate significance of the coverage probability at the 5% and 1% level, respectively, as measured by its asymptotic Normal distribution.

always undercovers, which means that the usual standard error bands around the point estimate are too tight.

Figure 4 shows the power curves of the Wald and AR tests of the hypothesis $H_0 : \varrho = \varrho_0$ at the 5% nominal level of significance for sample sizes $T = 100$ and 200. As we explain in the next section, the parameter ϱ measures the degree of indexation of prices to past inflation. It is evident that the AR test has good power, especially over the theoretically relevant parameter regions. In particular, it rejects with high probability the null hypothesis of $\varrho = 0$ (no indexation), when indexation is substantial, and thus, it can provide reliable evidence on this issue of considerable interest in applied work. The AR test does not have good power for high values of ϱ against higher alternatives. This means the test has difficulty distinguishing between a high degree and complete

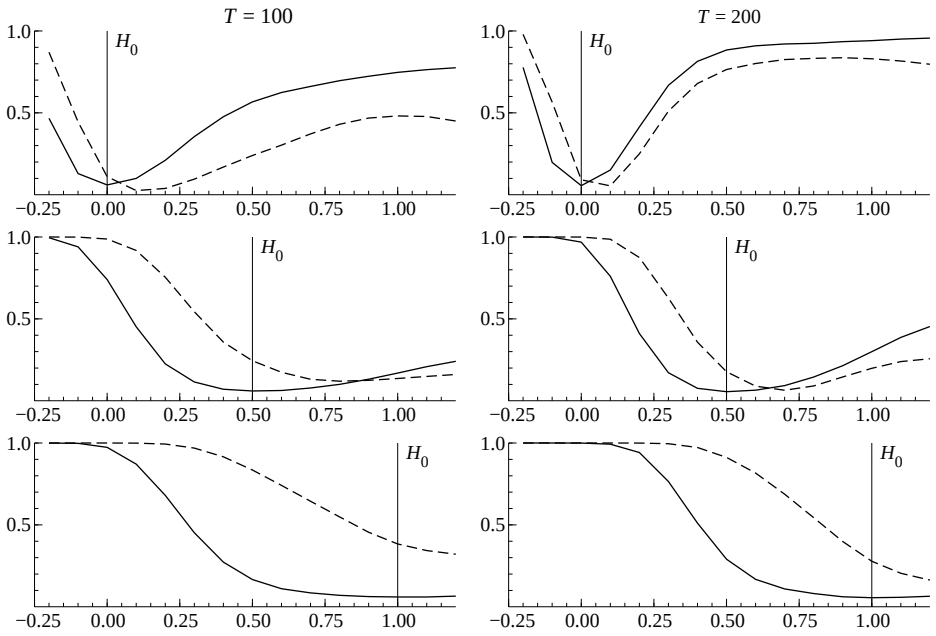

Figure 4: Power curves of the Wald (dotted line) and AR (solid line) tests for $T = 100$ (left column), and $T = 200$ (right column). The null hypothesis is $H_0 : \varrho = \varrho_0$, where $\varrho_0 = 0$ (top row), $\varrho_0 = 0.5$ (middle row), and $\varrho_0 = 1$ (bottom row), and it is superimposed by means of a vertical line. The value of ϱ under the alternative is shown by the abscissa. The other parameter values in the DGP are: $\beta = 0.99$, $\gamma = 0.01$, $\lambda = 0.15$, $\sigma_\varepsilon = 3$, $\sigma_{\varepsilon v} = 0.1$, $\rho_1 = 0.9$, $\rho_2 = 0$. The number of MC replications is $M = 10000$ and the nominal level of significance is 5%.

indexation.

Finally, as we discussed in section 2.1, higher values of the gain parameter, which are interpretable as a slower speed of learning, generate more variability of the adaptive forecasts relative to the corresponding RE forecast, and we expect this to have a positive impact on the accuracy of inference on the structural parameters. Figure 5 gives evidence of this effect through a direct comparison of the power function of the AR statistic at different values of γ and T. Specifically, the figure depicts the contour plots of the power function for the null hypothesis $H_0 : \varrho = 0$ against four alternatives, with respect to γ and T. It is clear that power increases in γ as well as T, and, moreover, that γ plays a role similar to the sample size, in that the same power can be achieved with a lower value of T and a higher value of γ.

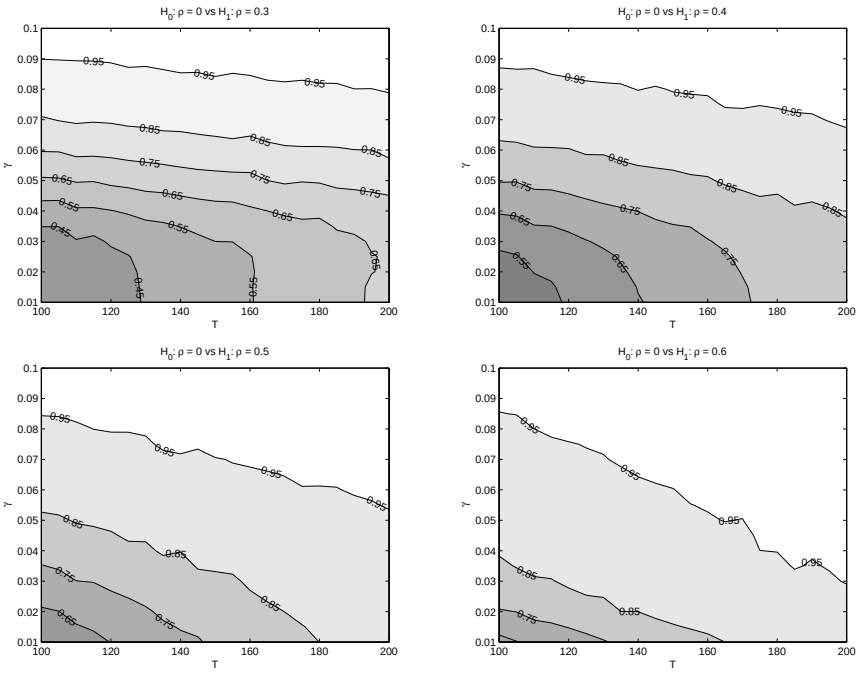

Figure 5: Power of the 5% level AR test of the null hypothesis $H_0 : \varrho = 0$ against the four alternatives: $\varrho = 0.3$, 0.4, 0.5 and 0.6. Each panel shows contours of the power function in terms of γ (the ordinate) and T (the abscissa). Power increases in the north-easterly direction in each of the four panels. The other parameter values are: $\beta = 0.99$, $\lambda = 0.15$, $\sigma_\varepsilon = 3$, $\sigma_{\varepsilon v} = 0.1$, $\rho_1 = 0.9$, $\rho_2 = 0$. The number of MC replications is $M = 10000$.

4 Application: the new Keynesian Phillips curve

4.1 Model and empirical issues

The NKPC is a purely forward-looking model of inflation dynamics which takes the form

$$\pi_t = \beta E_t \pi_{t+1} + \lambda \widehat{s}_t + u_t \tag{17}$$

where π_t denotes inflation, the forcing variable $\widehat{s}_t$ is a measure of real marginal costs in deviation from their steady state and u_t is an unobserved disturbance. The deep parameters of the model are the discount factor β and the degree of price stickiness ϑ, which is the probability that a firm will be unable to change its price in a given period. The slope of the Phillips curve λ is function of β and ϑ, $\lambda = (1-\vartheta)(1-\vartheta\beta)/\beta$.

Many studies report difficulties in fitting model (17) to US data when expectations are modelled as rational, see, for instance, Fuhrer and Moore (1995), Gali and Gertler (1999), Rudd and Whelan (2005, 2006). In particular, the model predicts that the dynamics of inflation should be explained solely by the dynamics of marginal costs, but this does not turn out to be the case with US data. In response to this empirical failure, the baseline purely forward-looking specification (17) has been extended to include lagged inflation on the right hand side. For example, assuming that a fraction ϱ of prices that cannot be reoptimized are indexed to past inflation, the model becomes

$$\pi_t = \beta E_t (\pi_{t+1} - \varrho \pi_t) + \varrho \pi_{t-1} + \lambda \widehat{s}_t + u_t. \tag{18}$$

see Woodford (2003, ch. 3 sec. 3.2) for details. Notably, the hybrid NKPC (18) nests the pure model (17) when $\varrho = 0$. The other polar case of complete indexation, $\varrho = 1$, is often used in empirical studies, see, e.g., Christiano, Eichenbaum, and Evans (2005).[9] Another common approach to introducing additional dynamics to the pure NKPC model (17) is to suppose that the unobservable cost push shock u_t is autocorrelated, as in Clarida, Gali, and Gertler (1999).[10] Yet another source of inflation persistence are time delays in the introduction of new prices. With a delay of d quarters,

[9] Galí and Gertler (1999) provided an alternative derivation of the NKPC, based on the idea that some fraction of firms set prices according to a backward-looking rule of thumb.. As Woodford (2003, p. 217) notes, the two models have identical implications in the limiting case $\beta = 1$.

[10] There are also studies that use both indexation and autocorrelated shocks, e.g., Smets and Wouters (2007).

the NKPC (18) becomes instead

$$\pi_t = \beta E_{t-d}\left(\pi_{t+1} - \varrho\pi_t\right) + \varrho\pi_{t-1} + \lambda E_{t-d}\widehat{s}_t + u_t \tag{19}$$

see Woodford (2003, p. 217).

Recently, Milani (2005, 2007) argued that if the assumption of rational expectations is replaced by some form of boundedly rational expectations, the pure NKPC model fits the data without the need to make potentially ad hoc assumptions to generate additional sources of persistence. Specifically, if inflation expectations are formed recursively by CGLS, they will depend on past data more than they would under rational expectations. Using OLS and Bayesian methods, Milani found that indexation is not statistically significantly different from zero.

Since our focus is primarily on the fit of the NKPC, we take a limited-information approach, following Galí and Gertler (1999) and Sbordone (2002). As Woodford (2003) explains, this approach makes weaker assumptions than full-information methods about the determinants of marginal costs, and is therefore more robust to misspecification of other parts of the system. Moreover, unlike Bayesian inference, our approach does not require the specification of the distribution of the shocks. Of course, weaker assumptions imply fewer identifying restrictions, and thus robustness comes at the cost of lower accuracy of inference. However, our results show, in line with the simulation evidence reported earlier, that our limited-information analysis is powerful enough to uncover new evidence on the empirical fit of the NKPC under learning, and our tests are highly informative about certain parameters of the model.

In our analysis we fix the discount factor β to 0.99. This assumption simplifies inference on the other three key parameters of the model ϱ, ϑ and the gain parameter, which are admittedly the more interesting ones. We note that our results remain robust if β is unrestricted.

Data Our estimation results are based on quarterly US data that cover the period 1960:Q2 to 2007:Q3. Following Galí and Gertler (1999) and Sbordone (2002), we derive our measure of $\widehat{s}_t$ assuming it is proportional to the log of the labor share.[11] The factor of proportionality depends on assumptions about factor markets and is typically calibrated. We set it to the value used by Galí and Gertler (1999), following the method of Sbordone (2002), so that our results are comparable to theirs.[12] Inflation is measured by the first difference in the logarithm of the (seasonally adjusted)

[11] We use the data reported by the Bureau of Labor Statistics (series ID: PRS85006173).

[12] Specifically, we set it to $(1-\omega\theta)^{-1}$ where θ is the elasticity of substitution between differentiated goods and $\omega = 1 - 1/\alpha_n$, where α_n is the elasticity of output with respect to labor in the production function. The parameters

implicit GDP deflator. We also use the Federal Funds rate as an additional instrument.

4.2 Results for the baseline specification

Our baseline specification is the hybrid NKPC (18) with expectations determined by perpetual learning, that is, CGLS. To close the model, we need to specify the PLM agents use to derive their forecasts. We assume the PLM is a vector autoregression of order p, VAR(p) in inflation and the labor share. This nests the simple specification of a first order autoregression for inflation, used, amongst others, in Milani (2005) and Orphanides and Williams (2005b). Also, under certain conditions on the law of motion of the labor share, it also nests the rational expectations equilibrium, as explained in Evans and Honkapohja (2001). This approach is also common in the literature, see Bullard and Eusepi (2005), Milani (2007) and Orphanides and Williams (2005a). The number of lags p in the VAR in our baseline specification is set to one based on information criteria, but we also investigate the robustness of the results to different choices of p.

Agents estimate the coefficients of the PLM recursively using the SRA (3)-(4) with constant gain parameter γ.[13] Their h-step ahead forecasts are then derived in the usual way.

Our identification assumption in the baseline model is that the disturbance term u_t is uncorrelated with its own lags and any other predetermined variables. This is identical to the assumption used in Galí and Gertler (1999), Sbordone (2002) and Milani (2005),[14] but it can be easily relaxed to allow for exogenous persistence in the cost push shock, e.g., $u_t = \rho_u u_{t-1} + \varepsilon_t$. We investigate the robustness of the results to this as well as more general alternatives.

We collect the parameters of interest in a vector $\theta = (\vartheta, \varrho, \gamma)'$ (recall that β is fixed to 0.99), and use the notation $\pi^e_{t+1}(\gamma)$ to denote explicitly the dependence of π^e_{t+1} on γ. We allow the parameters ϑ and ϱ to take values that are consistent with the underlying theory, namely $0 < \vartheta \leq 1$ and $0 \leq \varrho \leq 1$. In principle, γ could take any value between zero and one but we put an upper bound at 0.1. This is motivated by the assumptions in Milani (2007), but our results are robust to using

are calibrated such that the markup $\mu = \theta/(\theta - 1)$ is 1.1 (10%) and $\mu\alpha_n = 2/3$, following Gali and Gertler (1999).

[13] We initialize the learning algorithm by

$$R_{t_0} = \gamma \sum_{i=0}^{t_0-1} (1-\gamma)^j z_{t-i} z'_{t-i}, \quad a_{t_0} = R_{t_0}^{-1} \gamma \sum_{i=0}^{t_0-1} (1-\gamma)^j z_{t-i} y_t.$$

where in $y_t = (\pi_t, \widehat{s}_t)'$ and $z_t = (1, \pi_{t-1}, \ldots, \pi_{t-p}, \widehat{s}_{t-1}, \ldots, \widehat{s}_{t-p})$, using presample data. Carceles-Poveda and Giannitsarou (2007) discuss this and other alternative initialization schemes and show that the choice of initialization is unimportant for CGLS. Indeed, alternative initializations do not make material difference to our results.

[14] In their estimation, Galí and Gertler (1999) assumed rational expectations and replaced π^e_{t+1} by its realization π_{t+1}, thus causing the residual in the estimated model to be a moving average of order 1. This is consistent with the disturbance in the NKPC being serially uncorrelated.

higher upper bounds on γ.[15] We exclude zero, since π^e_{t+1} cannot be computed at zero using the CGLS algorithm, as explained in Carceles-Poveda and Giannitsarou (2007), though it can be argued that, provided $\beta < 1$, $\pi^e_{t+1}(\gamma)$ converges to the rational expectation of π_{t+1} as γ goes to zero, see Milani (2007) or Orphanides and Williams (2005b).

We compute the AR statistic at θ, $AR(\theta)$, using the formula (13) corrected for an unrestricted constant, where the residuals η_t are given by the following function of the data and the parameters:

$$h_t(\theta) = \pi_t - \beta\pi^e_{t+1|t}(\gamma) - \varrho(\pi_{t-1} - \beta\pi_t) - \frac{(1-\vartheta)(1-\beta\vartheta)}{\vartheta}\widehat{s}_t. \tag{20}$$

We employ four lags of the residuals $h_t(\theta)$ the labor share and the Fed Funds rate as instruments, and use White's (1980) heteroskedasticity consistent estimator for the variance in the AR statistic (13), in order to account for time-variation in the volatility of the shocks, given the evidence reported in the literature, see Sims and Zha (2006).

We construct $(1-\varphi)$-level confidence sets on the parameters θ by inverting a φ-level test based on the AR statistic. This is done by evaluating $AR(\theta)$ over the entire parameter space and collecting all the values of θ such that $AR(\theta)$ is less than the $1-\varphi$ quantile of the $\chi^2(k)$ distribution, where k is the number of instruments (twelve in the baseline model). If the confidence set is empty, it means that there is no θ for which the moment conditions are satisfied, i.e., the model is misspecified. Thus, the p-value (i.e., tail probability) corresponding to minimum value of $AR(\theta)$ serves as a measure and formal test of the model's fit. Moreover, the 'best-fitting value' $\widehat{\theta} = \arg\min_\theta AR(\theta)$ is the continuously updated GMM estimator (CUE) of θ, since $AR(\theta)$ can be interpreted as a continuously updated GMM objective function, see Hansen, Heaton, and Yaron (1996).

We start by assessing the fit of the baseline model (18). Our main finding is that the baseline model does not fit the data. The p-value associated with least rejected value of θ is 0.0002, indicating rejection at the 0.02% level of significance. The failure of the baseline model to match the data is robust to alternative specifications of π^e_{t+1}. Table 4 reports such robustness checks and shows that, whether π_t is included in the forecast of π_{t+1} or not, and whether higher order VAR specifications are used in the PLM, the p-value of the minimum AR statistic remains well below the 5% level of significance. As a further robustness check against potential misspecification of agents' forecasting model, we report the fit of the baseline model when π^e_{t+1} is measured using data from the Survey

[15] 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. Our results are robust to using an upper bound of 0.15 or 0.2.

Baseline specification: $\min AR(\theta) = 37.03$, p-value*: 0.0002

Alternative assumptions about π^e_{t+1}		
	$\min AR(\theta)$	p-value*
π_t included in π^e_{t+1}	31.11	0.002
PLM is VAR(2)	29.37	0.003
PLM is VAR(3)	24.63	0.017
PLM is VAR(4)	26.11	0.010

Alternative measures of π^e_{t+1}			
	$\min AR(\theta)$	p-value*	Available sample
Greenbook forecasts	41.65	0.000	1967q1-1995q4
Survey of prof. forecasters	58.80	0.000	1970q2-2007q3

*p-value is based on $\chi^2(12)$ distribution.

Table 4: Fit of the baseline NKPC with indexation

of Professional Forecasters, or real-time Greenbook data.[16] The results reported in Table 4 show that the baseline model remains resoundingly rejected.

A possible cause of misspecification is variation in the parameters of the model over time, such as a structural break. Parameter instability will result in a violation of the moment conditions of a model that incorrectly assumes the parameters are constant over the entire sample. Perhaps the simplest way to account for variation in the parameters is to check the fit of the model over subsamples. Looking at subsamples of ten years each, the results remain the same. In all of the subsamples the model is rejected at the 5% level.

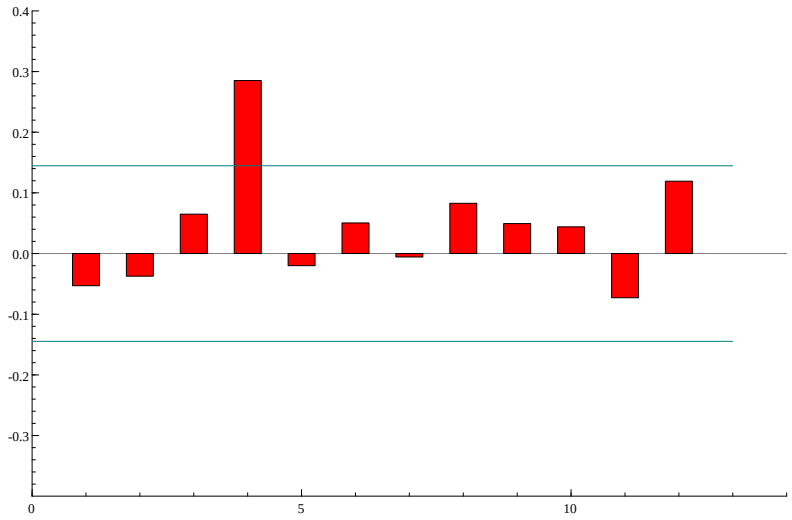

Figure 6: Correlogram of the residuals of the baseline NKPC model with indexation

[16] The Greenbook data were kindly provided by Orphanides, and are the data used in Orphanides (2004). The survey data are the median one-year-ahead forecasts of inflation compiled by the Philadelphia Fed.

p	$\min AR(\theta)$	p-value	d.f.
1	34.00	0.0004	11
2	32.38	0.0003	10
3	32.35	0.0002	9
4	12.45	0.1322	8

Shocks: $u_t \sim AR(p)$.
p-value is based on $\chi^2(df)$ dist.

Table 5: Fit of the NKPC with indexation and autocorrelated shocks

A look at the residuals of the model sheds light onto the possible sources of misspecification. Figure 6 plots the correlogram of the residuals up to twelve quarters. It is immediately apparent that the residuals exhibit significant autocorrelation at lag four that is not explained by the baseline specification of the model. Moreover, the structure of autocorrelation is such that it cannot be captured by modelling the residuals as AR(1). Table 5 reports tests of the fit of the model with alternative assumptions about the autocorrelation of the shock u_t in the NKPC (18). Consistently with the correlogram in Figure 6, the AR statistic remains significant at the 1% level unless u_t is modelled as AR(4). ARMA models for u_t may be used as alternatives to the autoregressive specification. However, it is clear from the picture that an ARMA(1,1) specification, as for example, in Smets and Wouters (2007), is not flexible enough to capture the serial correlation pattern of the residuals.[17] The type of model for u_t that is required to do so is not one that we have seen used in applied work. We therefore consider time delays in price changes as an alternative source of persistence in inflation, see equation (19), which is more appealing from a theoretical perspective than ad hoc assumptions about the autocorrelation of the errors.

4.3 Results for a model with time delays in price changes

We now turn to the NKPC with indexation and time delays of d quarters, which is given by equation (19). Preliminary estimates, reported in Table 7 of the appendix, show that this model is still unable to fit the data over the full sample. However, it appears that the model with $d = 4$ and AR(1) shocks fits the data when the parameters are allowed to be different over subperiods.

In fact, the only parameter that appears to vary significantly over time is the gain parameter. So, a model estimated over the entire sample that allows for changes only in the gain parameter suffices to fit the data. The gain parameter is interpretable as measuring the speed of learning,

[17] We are not suggesting that the version of the NKPC in Smets and Wouters (2007) suffers from misspecification of the kind that is discussed here, since their NKPC is different from our specification. We mention Smets and Wouters only as an exmple of an empirical study that models the disturbance in the NKPC as ARMA(1,1).

because higher gains are associated with faster discounting of past data in the estimation of the PLM. One may expect to see higher discounting of past observations over periods of instability, during which the use of constant parameter reduced-form models for forecasting is susceptible to the Lucas (1976) critique.

The above considerations motivate us to consider three periods of similar length across which the gain parameter may be different: 1960q1-1973q3, 1973q4-1987q3 and 1987q4-2007q3 The second period starts at the onset of the first oil price shock, covers the great inflation of the seventies and the subsequent disinflation of the early eighties, and ends in 1987q3, when Greenspan became the chairman of the Fed. The first and third periods are characterized by relative macroeconomic stability. The gain parameter is allowed to be different across periods, but constant within each period:

$$\gamma_t = \begin{cases} \gamma_1, & \text{before 1973q4} \\ \gamma_2, & \text{1973q4 to 1987q3} \\ \gamma_3, & \text{after 1987q3} \end{cases} \tag{21}$$

We estimate the NKPC (19) with indexation and time delay of four quarters, $d = 4$, and an AR(1) shock, allowing for a break in the gain parameter as in (21) – the model does not fit with $d < 4$ or serially uncorrelated shocks. Table 6 reports the estimation results. The p-value associated with $\min AR(\theta)$ is 0.3, indicating nonrejection at conventional significance levels. Thus, confidence sets on θ derived by inverting the AR statistic are non-empty at conventional levels. Figure 7 reports two-dimensional confidence sets on the two pairs of parameters (ϑ, ϱ) and (γ_1, γ_2). The shaded areas contain ϕ-level confidence sets for each pair, derived by the projection method, see Dufour and Taamouti (2005).

The point estimates reported in Table 6 suggest that the gain parameter is different across the three periods. Specifically, the gain parameter is high during the volatile period of 1973 to 1987, and is small before and after that period. In fact, the estimates of γ for the first and last period, $\widehat{\gamma}_1$ and $\widehat{\gamma}_3$ are identical, so we impose this restriction in the ensuing analysis. We can assess whether the difference in the gain parameters γ_1 and γ_2 is statistically significant using the AR statistic. The p-value associated with the hypothesis $\gamma_1 = \gamma_2 = \gamma_3$ is 0.03, showing that the change in the gain parameter is indeed statistically significant at the 5% level. This is also evident from the confidence set on (γ_1, γ_2) reported in the right panel of figure 7. Even though the confidence set is wide, showing that the gain parameters cannot be estimated very precisely, they are still informative about a break in the gain parameter, since the 95%-level confidence set does not cross

Parameter	Estimate	95% CI	90% CI
ϑ	0.65	$[0.38, 1]$	$[0.42, 1]$
ϱ	0.14	$[0, 0.46]$	$[0, 0.40]$
γ_1	0.01	$[0.01, 0.06]$	$[0.01, 0.05]$
γ_2	0.10	$[0.02, 0.1]$	$[0.03, 0.1]$
$\gamma_2 - \gamma_1$	0.09	$[0.01, 0.09]$	$[0.02, 0.09]$
$\min AR$: 12.93, p-value: 0.3 $[\chi^2(11)]$			

Table 6: The NKPC with time delays, autocorrelated shocks and structural change in the gain parameter: γ_2 is the value of the gain parameter in the period 1973q4-1987q3, and γ_1 is the value before and after that period.

the 45 degree line.

We now look at the estimates of the two structural parameters ϑ and ϱ. Consistently with Milani (2007), the indexation parameter ϱ is not significantly different from zero. Its point estimate is low (0.14) compared to other studies, but the confidence interval associated with it is wide, $[0, 0.4]$. Still, the confidence interval is small enough to reject a model with full indexation. The parameter governing the degree of price stickiness, ϑ, is notably very imprecisely estimated. Virtually all the parameter estimates reported in the literature fit within the 90% confidence interval $[0.42, 1]$.. Moreover, we cannot reject the null hypothesis that the Phillips curve is completely flat, $\vartheta = 1$.

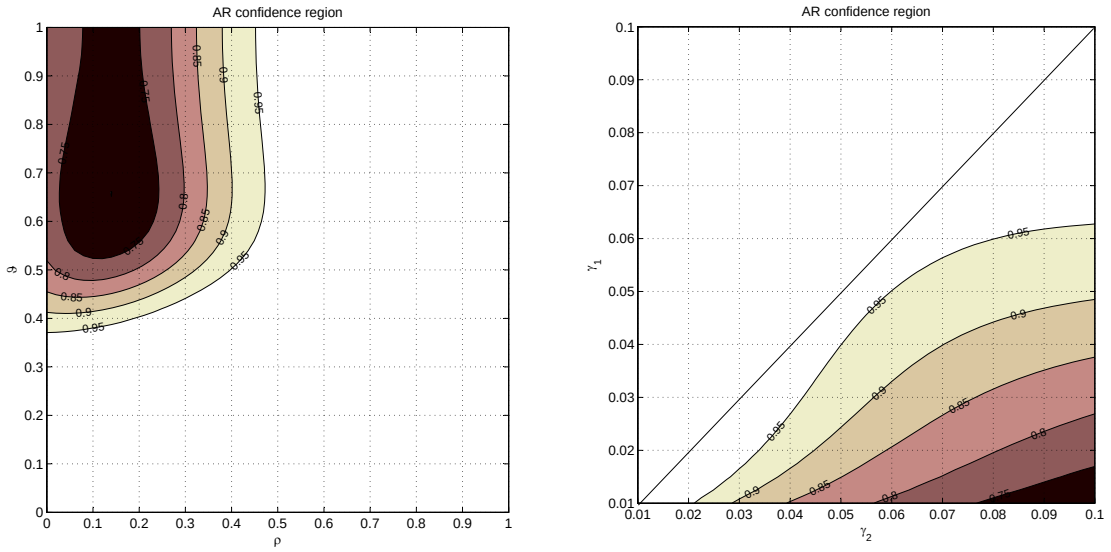

Figure 7: Two-dimensional confidence sets for (θ, ϱ) (left) and (γ_1, γ_2) (right) based on the Anderson Rubin statistic. The model is the NKPC with indexation, four-quarter time delay in price changes and AR(1) shock. Instruments include four lags of the shocks, the labor share and the Fed Funds rate, and White's HC estimator is used.

The above inference can be sharpened considerably by using more instruments, without altering

any of the conclusions reached above. When we use four additional lags of the residuals of the model as instruments, the confidence sets and intervals are generally tighter, and we can reject the null hypothesis of a constant gain across periods at the 0.1% level of significance.

4.4 Discussion

One important message of our analysis is that standard versions of the NKPC are unable to fit the dynamics in inflation, and this is as true under learning as it is under rational expectations, the latter shown, amongst others, by Rudd and Whelan (2005, 2006). In particular, the model fails to account for significant fourth order autocorrelation in inflation. To capture this feature of inflation dynamics, we considered a simple extension of the model that allows for time delays in price changes. There are certainly other ways of modelling dependence at the annual frequency. Time delays in price changes may be a relevant feature of most markets, but a delay of four quarters, which is needed for the NKPC to fit the data, may seem unrealistically long. It seems plausible that such dependence may be the result of wage contracts being negotiated on an annual basis, so, modelling wage and price setting behavior jointly may provide a more appealing explanation of this feature of the data.

Another important message of our analysis concerns the speed of learning. Unlike earlier work, we find evidence against a model of learning with small and constant gain. In particular, our results show that the gain parameter varies over time and is higher in periods of macroeconomic instability. Our attempt to model this time variation in the gain parameter is rather limited, of course, and is not intended as a structural alternative to CGLS. However, it is sufficient to provide reduced-form evidence against a constant gain specification of learning dynamics. Our empirical results suggest that it may be appropriate to make the gain parameter endogenous, as for instance in Marcet and Nicolini (2003). Such alternatives can be studied easily using the econometric method that we propose in this paper.

A Appendix

Proof of proposition 1. Solving equation (10) 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 β and γ using $\gamma = \psi/\sqrt{T}$ and $1-(1-\beta)\gamma = \exp(\phi/T)$, this can be written as:

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

As $T \to \infty$, then

$$T^{-1/2}\sum_{t=1}^{[Tr]} \eta_t \Rightarrow \sigma_\eta W(r)$$
$$\frac{\psi}{\sqrt{T}}\sum_{i=0}^{[Tr]-1} e^{\phi i/T}\eta_{[Tr]-i} \Rightarrow \psi\sigma_\eta J_\phi(r)$$

for $0 \le r \le 1$, see (Phillips, 1987, Lemma 1), where J_ϕ is an Ornstein-Uhlenbeck diffusion with $J_\phi(0) = 0$, and parameter ϕ, driven by the Brownian motion $W(r)$. Moreover, since $e^{\phi r} - e^{\phi t/T} \to 0$ as $T \to \infty$ uniformly in $0 \le r \le 1$, equation (11) follows by Slutsky's formula for weak convergence. Now, we turn to the OLS estimators:

$$\begin{bmatrix} \widehat{\beta}-\beta \\ \widehat{\delta}-\delta \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^T a_{t-1}^2 & \sum_{t=1}^T a_{t-1} \\ \sum_{t=1}^T a_{t-1} & T \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^T a_{t-1}\eta_t \\ \sum_{t=1}^T \eta_t \end{bmatrix}, \quad \text{or}$$
$$\begin{bmatrix} \sqrt{T}\left(\widehat{\beta}-\beta\right) \\ \sqrt{T}\left(\widehat{\delta}-\delta\right) \end{bmatrix} = \left(\begin{bmatrix} T^{-1}\sum_{t=1}^T a_{t-1}^2 & T^{-1}\sum_{t=1}^T a_{t-1} \\ T^{-1}\sum_{t=1}^T a_{t-1} & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} T^{-1/2}\sum_{t=1}^T a_{t-1}\eta_t \\ T^{-1/2}\sum_{t=1}^T \eta_t \end{bmatrix} \tag{22}$$

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 (12) follows. ■

shock:	serially uncorrelated		AR(1)	
d	$\min AR(\theta)$	p value	$\min AR(\theta)$	p value
1	33.20	0.0009	30.37	0.0014
2	42.77	0.0000	37.23	0.0001
3	41.25	0.0000	48.05	0.0000
4	34.45	0.0006	21.28	0.0306

Table 7: Fit of the NKPC with time delays

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.

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica 74*(3), 715–752.

Beyer, A. and R. E. A. Farmer (2007). Testing for indeterminacy: An application to U.S. monetary policy: Comment. *American Economic Review 97*(1), 524–529.

Bray, M. M. and N. E. Savin (1986). Rational expectations equilibria, learning, and model specification. *Econometrica 54*(5), 1129–1160.

Bullard, J. B. and S. Eusepi (2005). Did the great inflation occur despite policymaker commitment to a taylor rule? *Review of Economic Dynamics 8*, 3244–359.

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.

Christiano, L. J., M. Eichenbaum, and C. Evans (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *J. Political Economy 113*, 1–45.

Clarida, R., J. Gali, and M. Gertler (1999). The science of monetary policy: A new keynesian perspective. *Journal of Economic Literature 37*(4), 1661–1707.

Cochrane, J. H. (2007a). Identification with taylor rules: A critical review. NBER Working Papers 13410, National Bureau of Economic Research, Inc.

Cochrane, J. H. (2007b). Inflation determination with taylor rules: A critical review. NBER Working Papers 13409, National Bureau of Economic Research, Inc.

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.

Dufour, J.-M. and M. Taamouti (2005). Projection-based statistical inference in linear structural models with possibly weak instruments. *Econometrica 73*(4), 1351–1365.

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 rersearch. Discussion Paper 6640, CEPR.

Fourgeaud, C., C. Gourieroux, and J. Pradel (1986). Learning procedures and convergence to rationality. *Econometrica 54*(4), 845–68.

Galí, J. and M. Gertler (1999). Inflation dynamics: a structural econometric analysis. *Journal of Monetary Economics 44*, 195–222.

Gorodnichenko, Y. and S. Ng (2007). Estimation of dsge models when the data are persistent. Technical report. Preseented at NBER Summer Institute.

Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

Hansen, L. P., J. Heaton, and A. Yaron (1996). Finite sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics 14*, 262–280.

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. (2005). Testing parameters in GMM without assuming that they are identified. *Econometrica 73*(4), 1103–1123.

Lucas, R. E. (1973). Some international evidence on output-inflation tradeoffs. *American Economic Review 63*(3), 326–334.

Lucas, R. E. J. (1976). Econometric policy evaluation: a critique. In K. Brunner and A. Meltzer (Eds.), *The Philips Curve and Labor Markets.*, Carnegie-Rochester Conference Series on Public Policy. Amsterdam: North-Holland.

Marcet, A. and J. P. Nicolini (2003). Recurrent hyperinflations and inflation. *Americon Economic Review 93*, 1476–1498.

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. (2005). Adaptive learning and inflation persistence. Working Papers 050607, University of California-Irvine, Department of Economics.

Milani, F. (2007). Expectations, learning and macroeconomic persistence. *Journal of Monetary Economics 54*(7), 2065–2082.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica 71*(4), 1027–1048.

Nicholls, D. F. and A. R. Pagan (1983). Heteroscedasticity in models with lagged dependent variables. *Econometrica 51*(4), 1233–42.

Orphanides, A. (2004). Monetary policy rules, macroeconomic stability, and inflation: A view from the trenches. *Journal of Money, Credit and Banking 36*(2), 151–75.

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 (2005a). The decline of activist stabilization policy: Natural rate misperceptions, learning, and expectations. *Journal of Economic Dynamics and Control 29*(11), 1927–1950.

Orphanides, A. and J. C. Williams (2005b). Inflation scares and forecast-based monetary policy. *Review of Economic Dynamics 8*(2), 498–527.

Phillips, P. C. B. (1987). Towards a unified asymptotic theory for autoregression. *Biometrika 74*(3), 535–547.

Primiceri, G. E. (2006). Why inflation rose and fell: Policymakers' beliefs and us postwar stabilization policy. *Quarterly Journal of Economics 121*(3), 867–901.

Rudd, J. and K. Whelan (2005). New tests of the new-keynesian phillips curve. *Journal of Monetary Economics 52*(6), 1167–1181.

Rudd, J. and K. Whelan (2006). Can rational expectations sticky-price models explain inflation dynamics? *American Economic Review 96*(1), 303–320.

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.

Sims, C. A. and T. Zha (2006). Were there regime switches in u.s. monetary policy? *American Economic Review 96*(1), 54–81.

Smets, F. and R. Wouters (2007). Shocks and frictions in us business cycles: A bayesian dsge approach. *AER 97*(3), 586–606.

Staiger, D. and J. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*, 557–586.

Stock, J. H. and J. H. Wright (2000). GMM with weak identification. *Econometrica 68*(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.

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.