On the Initialization of Adaptive Learning in Macroeconomic Models

DOI:
10.1016/j.jedc.2017.03.002

Document Version
Accepted author manuscript

On the Initialization of Adaptive Learning in Macroeconomic Models*

MICHELE BERARDI
University of Manchester

JAQUESON K. GALIMBERTI†
ETH Zurich

Abstract

We review and evaluate methods previously adopted in the applied literature of adaptive learning in order to initialize agents' beliefs. Previous methods are classified into three broad classes: equilibrium-related, training sample-based, and estimation-based. We conduct several simulations comparing the accuracy of the initial estimates provided by these methods and how they affect the accuracy of other estimated model parameters. We find evidence against their joint estimation with standard moment conditions: as the accuracy of estimated initials tends to deteriorate with the sample size, spillover effects also deteriorate the accuracy of the estimates of the model's structural parameters. We show how this problem can be attenuated by penalizing the variance of estimation errors. Even so, the joint estimation of learning initials with other model parameters is still subject to severe distortions in small samples. We find that equilibrium-related and training sample-based initials are less prone to these issues. We also demonstrate the empirical relevance of our results by estimating a New Keynesian Phillips curve with learning, where we find that our estimation approach provides robustness to the initialization of learning. That allows us to conclude that under adaptive learning the degree of price stickiness is lower compared to inferences under rational expectations.

Keywords: expectations, adaptive learning, initialization, algorithms, hybrid New Keynesian Phillips curve.

JEL codes: C63, D84, E03, E37.

1 Introduction

Adaptive learning algorithms have been proposed to provide a procedural rationality view on agents' process of expectations formation. Reopening a long standing debate on how should

*We gratefully acknowledge the comments and evaluations provided by an Associate Editor and two anonymous reviewers. An earlier version of this paper was presented at the 2016 EEA-ESEM joint meeting in Geneva. Any remaining errors are our own.

†Corresponding author. E-mail: galimberti@kof.ethz.ch.

expectations be modeled in macroeconomic models, the heuristics provided by learning algorithms come at the cost of introducing new degrees of freedom into the analysis. One such an open node relates to how these recursive mechanisms should be initialized in order to be representative of agents' learning-to-forecast behavior.

The main characteristic of the adaptive learning approach is its reliance on recursive algorithms in order to represent how agents update their beliefs as new observations about the economic relationship of interest become available. Such recursions naturally demand an initial starting point, and it is the numerical specification of these conditions that we denote as the initialization problem. Clearly, the uncertainties affecting the initialization of the learning process will propagate recursively into the predictions obtained with the model, and it seems crucial that the researcher understands the magnitude of these distortions and how they can affect structural inferences.

In this paper we investigate this issue with particular attention to the applied literature of learning in macroeconomics. Here applied is taken to encompass both theoretical simulations as well as exercises of empirical estimation and calibration. Examples can be found in Sargent (1999); Marcet and Nicolini (2003), or more recently in Eusepi and Preston (2011); Milani (2011), between many others cited throughout the paper. The main distinctive feature of these works consists in the replacement of the rational expectations (RE) assumption of an instantaneous adjustment of agents expectations, with a characterization of agents as adaptive learners of their own environment. More generally, our study will be relevant for scholars interested in the methods needed to uncover the initial beliefs of economic agents in models where such beliefs actually matter for economic dynamics.

The economic relevance of the initialization issue can be illustrated considering the long debated causes of the period of Great Inflation during the 1970s in the US. One of the main explanations for that episode comes from Sargent's (1999) hypothesis that the evolution of US inflation rates over the period can be attributed to the evolution of the monetary authority's beliefs about the trade-off between inflation and unemployment, the so-called Phillips curve. Subsequent studies advanced on this issue attributing the rise in inflation rates to delayed policy responses to ongoing structural changes in the economy of that period (see, e.g., Bullard and Eusepi, 2005; Orphanides and Williams, 2005a). Importantly, as evidenced in Primiceri (2006); Sargent et al. (2006), the point of departure in policymaker's beliefs, which is what we refer to as the learning initials in this particular context, is a crucial feature in such explanation.

Assumptions about initial beliefs also matter for the fit of models that introduce adaptive learning on the side of other market participants, such as households and firms. Examples are given by Carceles-Poveda and Giannitsarou (2008) for asset pricing models, Huang et al. (2009) in a standard growth model, and Slobodyan and Wouters (2012a,b) in a medium-scale dynamic stochastic general equilibrium (DSGE) model. Overall, these studies present results showing that whereas the introduction of learning has interesting effects on the dynamics and the fit of models to the data, a great portion of the improvements may be associated to transition

dynamics from specific initial beliefs. Hence, it is important to have a systematic evaluation of the different alternatives available as initialization methods, and we attempt to fill that gap in this paper.

We review the literature in order to pool together the pre-existing initialization methods into an archetypal classification that can be broadly defined in three major classes: equilibrium-related methods, training sample-based methods, and estimation-based methods. The equilibrium-related initializations are generally obtained taking rational expectations equilibrium (REE) as a reference, and exploring distributional deviations from that assumption. The training sample-based initializations, as the name suggests, are obtained with the application of the learning algorithm (or variations) over a pre-sample of observations that is left aside from the original sample of data available. Here we distinguish between two main forms of this method: (i) a Weighted Least Squares (WLS) approach, which is equivalent to the application of a constant gain Least Squares (LS) algorithm[1] to the training sample; and, (ii) an Ordinary Least Squares (OLS) approach, which is equivalent to the decreasing gain form of the LS algorithm. Finally, the more recent estimation-based methods consist of approaches involving the joint estimation of the initials with other model parameters, hence allowing the use of the same data that is used for inferences about structural features of the model to guide the specification of learning initials.

We compare the initialization methods on the basis of the accuracy of their delivered initial estimates and their effects over the accuracy of other estimated model parameters. To evaluate accuracy we derive measures of the Mean Squared Deviation (MSD) between true parameter values and their corresponding estimates obtained according to the different initialization methods. We relate the MSD measures to two main principles to judge the quality of an initialization. First, we look at the coherence of the initialization estimates to the dynamics implied by the learning process; second, we consider how susceptible the method is to biases that push up the model's explanatory power over the initial portion of observations in the sample.

The accuracy of initialization methods is first examined analytically within the simplified framework of an example model where agents are required to learn only a constant. Under these conditions we establish some important relationships between the invariant distribution of the learning estimates and the MSDs associated to the initialization methods: (i) an REE-based initial yields an MSD equal to the variance of the learning estimates; (ii) the accuracy of training sample-based initials depends on the number of observations left aside for training, and the specification of the learning gains; under a constant gain equal to that underlying the data generating process (WLS-based), MSDs tend to zero as the training sample increases; under decreasing gains (OLS-based), the MSDs converge to those obtained under the REE-based initial; (iii) an initial estimate based on the minimization of the model's sum of squared residuals tends to yield smaller MSDs as the estimation sample grows. However, we show that

[1]Consistent with the majority of the adaptive learning literature, in this paper we focus on the LS algorithm as representative of how agents update their expectations (see also Berardi and Galimberti, 2014, on this point).

only the WLS-based method is capable of converging to the true initial estimate. We believe these results should prove useful for applied researchers interested in the simulation of models with adaptive learning.

For empirical purposes, one is confronted with a more intricate initialization problem, namely, the estimation of other model parameters. To evaluate initialization methods under these circumstances we conduct several simulation exercises based on a model of the New Keynesian Phillips curve (NKPC) with learning. Our simulation results point to some interesting findings. First, the analytical results based on the assumption of known model parameters are confirmed for the NKPC model; particularly, the training sample-based initialization methods are in general favored in terms of the coherence criterion, since these methods are found to provide more accurate estimates of the learning initials. Second, the performance of the initialization methods is sensitive to the accuracy of their associated model parameters' estimates; here the equilibrium-related initials can prove useful for being less sensitive to the estimation of the learning gain, which turns out to be severely affected by finite sample distortions.

Notwithstanding, we find that the pre-determined initials were much less vulnerable to the estimation of other model parameters than the jointly estimated initials. Regarding the latter, we find that the estimation approach plays a key role in determining the quality of the initials and their effects over other parameters' estimates. Particularly, an estimation based solely on traditional moment conditions, derived from the assumption that the model's unobserved disturbance is a martingale difference sequence, leads to severe distortions to the estimated initials as the sample size grows. Furthermore, these distortions can spillover to the estimates of the other model parameters. Based on our analytical derivations, we proposed a solution to this problem with the inclusion of a squared residual criterion to the estimation objective, denoting this as the augmented approach. Although we show that our augmented approach is successful in restoring the consistency of the jointly estimated initials, this solution is only effective for large samples. Importantly, under the pre-determined initials the negative spillover effect from the initialization errors to the model parameters' estimates is not as strong as observed under the joint estimation approaches.

This last finding is of particular relevance for empirical analysis, where interest is usually in uncovering the underlying values of structural parameters that may validate the model's consistency with data evidence. In order to further enhance our understanding on the relevance of these different initialization methods for applied macroeconomics, we also present an empirical application on the determination of US inflation rates under the Phillips curve framework. Adopting a generalized method of moments (GMM) estimation approach, we find that the initials and the estimation criterion can, indeed, affect the estimates of structural parameters. Our results indicate that allowing for adaptive learning in the determination of inflation rates leads to a lower degree of price stickiness. This finding, nevertheless, is only robust to the alternative specifications of the initial estimates under our augmented estimation approach.

Finally, we also contribute to the literature on the estimation of models with adaptive learning by proposing the introduction of robust stability restrictions in the estimation of the model. Because adaptive learning under constant gain can generate unrealistic dynamics due to instabilities in the recursive estimation of the learning coefficients (see Evans and Honkapohja, 2009), the estimation of models under learning can be extremely sensitive to the sample of observations and the range of values allowed for the learning gain. To deal with this issue we impose additional constraints to the estimates of the learning gains, drawing upper bounds conditioned on the values of other parameters through simulations of the model. One advantage of this approach, compared to an unconditional reduction of the gain upper bound, is that it provides weaker constraints robust to learning stability for the joint estimation of the model parameters with the learning gain and initials.

The remainder of this paper proceeds as follows. In section §2 we provide a brief introduction to the use of adaptive learning in macroeconomic models, and establish the initialization problem under a simple "learn-the-average" example model. A review of the initialization methods previously adopted in the literature is presented in section §3 together with an analytical evaluation of these methods under the example model. We then proceed to present our simulation analysis, in section §4, and an empirical application, in section §5, both aiming at a comparative evaluation between the different methods of initialization under more realistic circumstances, such as the case where the learning gain and other model parameters require estimation. Finally, we conclude this paper with some remarks in section §6. Some key derivations and supplementary results are provided in the Appendices.

2 Adaptive Learning and the Initialization Problem

2.1 A brief primer on adaptive learning

Adaptive learning is introduced in macroeconomic models as an alternative to the assumption that agents hold rational expectations. One implication of the rational expectations assumption is that agents' beliefs are always consistent with the true model of the economy. Hence, under RE the economy instantaneously adjusts itself towards an equilibrium after any kind of shock that may have realistically affected agents' beliefs. In contrast, adaptive learning introduces some degree of persistence in the process through which agents update their beliefs, which allows such beliefs to deviate from RE in the short run, while keeping up with the idea of consistency in the long run.

To help fix ideas consider an univariate linear forward-looking model, where the determination of the current value of a variable of interest, y_t, depends on the value expected for that same variable in the next period, y^e_{t+1} plus a mean zero random shock, u_t, i.e.,

$$y_t = \beta y^e_{t+1} + u_t. \tag{1}$$

Simple as it stands, this specification may represent the reduced form of the equilibrium equations of an economic model which could potentially be non linear; it also corresponds, e.g., to simplified versions of two well known models: the Cagan model of inflation, letting y_t stand for the price level and u_t for a mean zero random supply of money; and, the standard model of asset pricing under risk neutrality, letting y_t stand for the asset price and u_t for a mean zero random sequence of dividends.

A solution to model (1) requires the specification of agents' perceived law of motion (PLM), which depicts how agents form expectations. Particularly, if agents condition their forecasts on a constant, $y^e_{t+1} = 0$ solves the model for any β. Hence, the stochastic process followed by the economy, also known as the actual law of motion (ALM), is directly determined by the specification of agents' PLM.

Under learning the corresponding ALM is given by

$$y_t = \beta\phi_{t-1} + u_t, \tag{2}$$

where ϕ_{t-1} denotes agents' estimates of the constant in their PLM based on observations available up to the previous period. Different recursive algorithms have been proposed in the literature to represent how agents update such estimates. Due to its widespread popularity between econometricians, one natural choice for that purpose has been the Least Squares (LS) algorithm, which can be generally defined as follows.

Algorithm 1 (LS)**.** *Let agents' PLM of y_t be given by a linear regression of the form*

$$y_t = \mathbf{x}_t'\boldsymbol{\phi}_{t-1} + \varepsilon_t, \tag{3}$$

where $\mathbf{x}_t = (x_{1,t}, \ldots, x_{K,t})'$ is a set of pre-determined variables, possibly including a constant (e.g., $x_{1,t} = 1$) and lags of y_t, $\boldsymbol{\phi}_t = (\phi_{1,t}, \ldots, \phi_{K,t})'$ stands for a vector of coefficients, possibly time-varying, and ε_t denotes an unpredictable disturbance term. Under this context, the LS estimates of $\boldsymbol{\phi}_t$, conditional on observations up to time t, are given by

$$\hat{\boldsymbol{\phi}}_t = \hat{\boldsymbol{\phi}}_{t-1} + \gamma_t\mathbf{R}_t^{-1}\mathbf{x}_t\left(y_t - \mathbf{x}_t'\hat{\boldsymbol{\phi}}_{t-1}\right), \tag{4}$$

$$\mathbf{R}_t = \mathbf{R}_{t-1} + \gamma_t\left(\mathbf{x}_t\mathbf{x}_t' - \mathbf{R}_{t-1}\right), \tag{5}$$

where γ_t is a learning gain parameter, and $\mathbf{R}_t$ stands for an estimate of the regressors matrix of second moments.

The LS algorithm is originally motivated as the result from the minimization of a weighted sum of squared errors, where the weights are determined by the learning gain parameter (see Berardi and Galimberti, 2013). Hence, the learning gain stands for a parameter determining how quickly a given information is incorporated into the algorithm's coefficients estimates. There are two particular cases of interest: (i) when $\gamma_t = 1/t$, every observation receives the

same weight and (4)-(5) reduces to the (recursive) Ordinary Least Squares (OLS); and, (ii) under a constant gain, past observations receive geometrically decaying weights and (4)-(5) can be viewed as a (recursive) Weighted Least Squares (WLS) with weights given by $(1-\gamma)^j$, where j indexes for the number of periods between the weighted observation and the last observation in the sample. Our focus is on the constant gain specification due to its relevance for applied purposes: because it allows for a continuous operation of the algorithm's tracking capabilities, the constant gain can capture time-varying effects of different sources, such as structural breaks or the out-of-equilibrium dynamics generated by stochastic shocks.

For a PLM with an intercept only the constant-gain LS algorithm simplifies to

$$\phi_t = \phi_{t-1} + \gamma\left(y_t - \phi_{t-1}\right). \tag{6}$$

Substituting y_t from model (1)'s ALM one can find that

$$\phi_t = \delta^t \phi_0 + \gamma \sum_{i=0}^{t-1} \delta^i u_{t-i}, \tag{7}$$

where $\delta = 1 - \gamma(1-\beta)$. Taking the unconditional expectation of this expression we find that convergence to the RE equilibrium requires that $|\delta| < 1$, which is only possible when $\beta < 1$ and $0 < \gamma < 2/(1-\beta)$. Notice this last condition is always true for $0 < \beta < 1$ and meaningful gain values in the interval $0 < \gamma < 1$. Under this assumption the variance of the long run distribution of the learning estimates around the REE, denoted by $\overline{\sigma}^2_\phi$, is then given by

$$\begin{aligned} \overline{\sigma}^2_\phi &= \lim_{t\to\infty} E\left[\phi_t^2\right], \\ &= \frac{\gamma\sigma_u^2}{(1-\beta)(1+\delta)}, \end{aligned} \tag{8}$$

where σ_u^2 is the variance of u_t. Thus, the dispersion of the learning estimates around the REE increases with the value of the learning gain; see Evans and Honkapohja (2001, Theorem 7.9) for a more general result on the relationship between the learning gain and the invariant distribution of the learning estimates.

2.2 The initialization problem

Recursive learning algorithms naturally demand an initial starting point, and it is the numerical specification of these conditions that we denote as the initialization problem. By the recursive nature of learning, any error in the initial estimates will propagate recursively into the predictions obtained with the model. Consider the case of our example model, (2)+(6); letting $\hat{\phi}_0$ stand for a guess of the true value of ϕ_0, the model prediction of y_{t+1} associated with this initial

is given by $\hat{y}_{t+1} = \beta\hat{\phi}_t$, where $\hat{\phi}_t$ is obtained from (4) as

$$\begin{aligned} \phi_0 &= \hat{\phi}_0, \\ \hat{\phi}_1 &= \hat{\phi}_0 + \gamma\left(y_1 - \hat{\phi}_0\right), \\ \vdots &\quad \vdots \\ \hat{\phi}_t &= \hat{\phi}_{t-1} + \gamma\left(y_t - \hat{\phi}_{t-1}\right). \end{aligned} \tag{9}$$

Let the corresponding prediction error be denoted by $\hat{\Delta}_{t+1} = y_{t+1} - \hat{y}_{t+1}$; then, the mean squared prediction error (MSPE) from this model amounts to

$$\begin{aligned} E\left[\hat{\Delta}_{t+1}^2\right] &= E\left[\left(\beta\left(\phi_t - \hat{\phi}_t\right) + u_{t+1}\right)^2\right], \\ &= \beta^2\lambda^{2t}\mathcal{D}_{\phi_0} + \sigma_u^2, \end{aligned} \tag{10}$$

where $\lambda = 1-\gamma$, and $\mathcal{D}_{\phi_0}$ stands for the initialization's Mean Squared Deviation (MSD), which is more generally defined as:

Definition 1 (MSD)**.** The Mean Squared Deviation between the true value of a parameter, θ_t, which may be a learning initial (e.g., ϕ_0, R_0) or a time-invariant model parameter, and a corresponding estimate, $\hat{\theta}_t$, is given by

$$\mathcal{D}_{\theta_t} = E\left[\left(\theta_t - \hat{\theta}_t\right)^2\right]. \tag{11}$$

Clearly, assuming that $0 < \gamma < 1$, (10) shows that the effects of initialization errors tend to disappear as the distance from the initial point increases. Also notice that as γ increases, the prediction error associated to an initialization error decreases. Hence, the smaller the learning gain, the more important are the learning initials for the accuracy of the predictions obtained with the model. For empirical purposes both the model and learning parameters, say β and γ in our example model, are not known *a priori* and may therefore require estimation. In that case initialization errors can further aggravate the accuracy of the model predictions through its effects on parameters' estimation error. In fact, as we will show in our simulation analysis further below, the estimation of the model and learning parameters can be severely affected by the misspecification of the learning initials.

2.3 Evaluation criteria

Our relative assessment of the initialization methods that we describe in the next section will be guided by two principles that we consider of relevance for applied adaptive learning research: (i) the initials COHERENCE to the learning process; and, (ii) the initials SUSCEPTIBILITY to bias the model's explanatory power and the estimation of its parameters.

In empirical settings, a proper initialization of the learning algorithm requires to find out what were agents' beliefs at the beginning of the sample of data. To achieve this goal it is important to understand the statistical properties of the learning process we are trying to mimic. Recursive estimation algorithms are statistically characterized by two main distinct phases: a transient phase, where the estimates are so far apart from the true parameter values that the upcoming sequence of updates can easily achieve substantial improvements to the accuracy of the estimates; and a stationary phase, where most of the updates to the estimates are essentially just tracking tiny disturbances that may affect the system under estimation. Hence, if the initial beliefs should reflect the continuation of an estimation process that was already in motion prior to the sample beginning, an initialization method will satisfy the COHERENCE criterion when it can deliver estimates as close as possible to the algorithm's long run operation[2]. Under our evaluation measure this corresponds to a minimization of the initials' MSD.

Another empirical issue is how much can the learning initials affect the accuracy of the estimates of other model parameters. Under standard likelihood-based estimation approaches, every data point is given the same weight on the estimation of a structural parameter that is assumed to be constant throughout the sample period. Under learning this weighting profile can be easily manipulated by tweaking the initial learning estimates so as to induce a transient phase in the portion of the sample that follows the initialization, which potentially increases the model's explanatory power. In the context of our analysis, such SUSCEPTIBILITY to biases is measured by looking at the MSDs of some key parameter estimates across the different initialization methods.

These principles may inherently generate a trade-off for the assessment of the initialization methods: one can always give up some degree of the COHERENCE delivered by a learning initial in exchange for some SUSCEPTIBILITY to tweak that initial in order to improve the model fit to the data. Thus, it is important to obtain a quantitative assessment of how the different methods perform with respect to these criteria.

3 Review of Initialization Methods

In this section we review initialization methods adopted in the previous literature. We also present expressions for their associated MSD under the example model described in the previous section, i.e., (2)+(6). For that purpose, throughout this section we assumed that: (i) both the learning gain and the model parameter are known; and, (ii) the learning process has already converged to its invariant distribution at the period where the initial is evaluated, and at the beginning of the training sample, when applicable. Numerical simulations validating our results are presented in Appendix C.1.

[2]Admittedly, one may also be interested in obtaining initials that reflect the transient phase that follows the occurrence of a large shock that shifted agents' beliefs away from equilibrium just before the initialization data point.

3.1 Equilibrium-related methods

One way to initialize learning algorithms is obtained by using the existing knowledge about the law of motion(s) generating the data. Particularly, conditional on the knowledge about the model specification and the parameter values, one can easily obtain the REE implied values of agents' PLM coefficients and use these equilibrium values as reference for the initial estimates. In the case of our example model, such initial would be given by $\hat{\phi}_0^{REE} = 0$, and the associated MSD equals to the variance of the learning estimates, $\overline{\sigma}_\phi^2$, as given by (8).

Although this method was naturally appealing in earlier works with theoretical simulations, such as in Bray and Savin (1986), its debut into the applied literature came in the seminal contribution of Sargent (1999). Its usage has since been prominent in studies on the effects of replacing the assumption of frictionless REE by the sticky process of expectations formation through adaptive learning (e.g., Marcet and Nicolini, 2003; Bullard and Eusepi, 2005; Orphanides and Williams, 2005b). For simulations, robust inferences can be obtained through this method by drawing the initials from a distribution centered around the REE values (see Carceles-Poveda and Giannitsarou, 2007).

Empirically, the uncertainties about the true model parameters may complicate the adoption of this method. One alternative is to approach the issue in two stages: first, model estimates are obtained under the REE assumption; second, these estimates are used to calculate the PLM coefficient values corresponding to the REE, which are then plugged back in as initial estimates for the algorithm's recursion for the analysis under learning (see Slobodyan and Wouters, 2012b; Ormeño and Molnár, 2015). One criticism to this practical solution is that it seems very likely that the REE estimates obtained in the first step will be biased for not taking the learning effects into account. Slobodyan and Wouters (2012a, p. 93) relax the first step by using model estimates obtained under a baseline learning specification instead of under RE, though still fixing the implied initials in the second step. Later, in our simulation and empirical exercises, we show that this issue can be circumvented by allowing the REE initials to be determined jointly with the model parameters under learning.

The REE-based initials do not provide ideal initial estimates for cases where there is prior information that the economy was in a transient phase at the beginning of the sample. In such a case, an alternative is provided by the *ad-hoc* initialization method, where the initials are hand-picked by the researcher. When taking the REE-based initials as a reference, this method provides a way to validate the sensitivity of results obtained under the former approach (e.g., Milani, 2007; Carceles-Poveda and Giannitsarou, 2008). In fact, one of the main uses of *ad-hoc* initials is to deal with the possibility of structural changes around the periods of the initials: when the changes affect the REE, agents may not be able to instantaneously adjust to the new equilibrium, and could therefore be forming expectations consistent with the previous equilibrium at the time of the initialization (see also Carceles-Poveda and Giannitsarou, 2007, p. 2679).

3.2 Training sample-based methods

Another common approach in learning applications is to use a pre-sample of observations in order to obtain the initial estimates. This is especially recommended for the cases where there is not enough previous knowledge about the system under estimation so as to allow an educated guess. The origins of this approach can be traced back into the engineering literature (see, e.g., Ljung and Soderstrom, 1983, pp. 299-303), where it is often suggested that the coefficients should be initialized with the value of zero and an initial training sample should be left aside to let the algorithm adjust its estimates according to the underlying calibration.

For applied purposes, it is often easier to adopt the non-recursive version of the learning algorithm to estimate the initials over the training sample. Letting P denote the number of observations set aside for the initialization, application of (4)-(5) in the training sample results in

$$\mathbf{R}_P = \gamma_P \sum_{i=1}^{P} w_i \mathbf{x}_i \mathbf{x}_i' + w_0 \mathbf{R}_\emptyset, \tag{12}$$

$$\hat{\boldsymbol{\phi}}_P = \gamma_P \mathbf{R}_P^{-1} \sum_{i=1}^{P} w_i \mathbf{x}_i y_i, \tag{13}$$

where $\{w_i\}_{i=0}^{P}$ stands for the sequence of weights given to each observation in the training sample, and $\mathbf{R}_\emptyset$ may incorporate prior information regarding the uncertainty surrounding the determination of the coefficients estimates. Under the assumption of a Gaussian random walk parameter drift model for $\boldsymbol{\phi}_t$, Berardi and Galimberti (2013) have shown that $\mathbf{R}_t$ is inversely related to the matrix of mean squared errors associated to the Kalman filter coefficients estimates, $E\left[\left(\boldsymbol{\phi}_t - \hat{\boldsymbol{\phi}}_t\right)\left(\boldsymbol{\phi}_t - \hat{\boldsymbol{\phi}}_t\right)'\right]$. Hence, in a Bayesian interpretation, as $\mathbf{R}_\emptyset \to \mathbf{0}$ the prior becomes more diffuse, since it is associated with a higher uncertainty about the coefficients estimates[3].

Depending on the weighting scheme and the prior estimates, there are two main variations of this method in the learning literature: the OLS-based (e.g., Williams, 2003; Orphanides and Williams, 2005a; Sargent et al., 2006) and the WLS-based (e.g., Primiceri, 2006; Milani, 2007, 2008, 2011, 2014; Huang et al., 2009; Chevillon et al., 2010; Eusepi and Preston, 2011; Lubik and Matthes, 2014) initials.

3.2.1 WLS-based initials

The WLS-based method derives from the Weighted Least Squares interpretation of the learning algorithm under a constant gain specification. In the training sample the initials associated to this method are obtained by setting $\gamma_P = \gamma$, and $w_i = (1-\gamma)^{P-i}$ in (12)-(13). Under our

[3]Notice that when $\mathbf{R}_\emptyset = \mathbf{0}$, a necessary condition for $\mathbf{R}_P$ to be invertible, as required in (13), is that $P \geq K$.

example model, the MSD associated to the WLS initial is given by

$$\begin{aligned}\mathcal{D}_{\phi_0}^{WLS} &= E\left[\left(\lambda^P \phi_{-P}\right)^2\right], \\ &= \lambda^{2P}\overline{\sigma}_{\phi}^2, \end{aligned} \tag{14}$$

from which we obtain the following proposition.

Proposition 1. *Under the model given by (2)+(6), the WLS-based initialization yields, on average, smaller MSDs than the REE-based initials.*

Remark 1. This result is specific to the model under consideration, and it stems from the fact that the REE is equal to the diffuse prior used for the WLS initial, i.e., $\phi_{REE} = \hat{\phi}_{-P}^{WLS} = 0$. Because the WLS method incorporates information from the training sample to that prior, it will always, on average, attain more precise initial estimates. In fact, if we add a constant α to the model in (1), $\phi_{REE} = \alpha/\left(1-\beta\right)$ and $\lim_{t\to\infty} E\left[\phi_t^2\right] = \overline{\sigma}_{\phi}^2 + \phi_{REE}^2$; then, whereas $\mathcal{D}_{\phi_0}^{REE}$ would still be equal to the variance of the invariant distribution of learning estimates, $\mathcal{D}^{WLS}$ would carry the effect of ϕ_{REE}^2.

For models with regressors in the PLM, as we will see in the simulation exercise of next section, the initialization of the matrix of second moments, (5), also becomes important. Here we consider two possibilities for the prior, $\mathbf{R}_{\emptyset}$. One, based on REE reasoning, is to set it to the sample estimate of the long-run covariance matrix of the regressors. Ideally, the sample used for such estimation should be restricted to the training sample itself, in order to prevent contamination of the initials due to the effects of changes in the statistical properties of the data that were not present before the initialization period. The second alternative we consider is to follow a diffuse approach and set $\mathbf{R}_{\emptyset} = \mathbf{0}$.

3.2.2 OLS-based initials

The OLS-based method, as the abbreviation suggests, is based on the Ordinary Least Squares estimator that is widely known among econometricians for possessing some well desired properties, such as consistency and efficiency in the estimation of linear models. For training sample initialization, it is obtained by setting $\gamma_P = 1/P$, $w_i = 1$, and $\mathbf{R}_{\emptyset} = 0$ in (12)-(13). One important advantage of this method relates to its convergence speed: the fact that a relatively higher gain value is used in the first iterations of the algorithm within the training sample tends to accelerate its convergence to the true initials.

Under our example model, the MSD associated to the OLS-based initial is given by

$$\mathcal{D}_{\phi_0}^{OLS} = \left(\delta^P - \frac{\beta\left(\delta^P - 1\right)}{P\left(\delta - 1\right)}\right)^2 \overline{\sigma}_{\phi}^2 + \sigma_u^2 \mathcal{K}, \tag{15}$$

where $\mathcal{K}$ is a function of the model parameters and the length of the initialization sample. Clearly, a necessary, though not sufficient, condition for the OLS-based initials to yield smaller MSD than the WLS-based initial is that $\lambda^P > \delta^P - \beta\left(\delta^P - 1\right) P^{-1} \left(\delta - 1\right)^{-1}$. However, notice that as $P \to \infty$, $\mathcal{D}_{\phi_0}^{OLS} \to E\left[\phi_t^2\right]$, which leads to our next proposition.

Proposition 2. *Under the model given by (2)+(6), as the length of the training sample increases to infinity, the MSD of the OLS-based initialization converges, on average, to the MSD of the REE-based initialization.*

Remark 2. A direct implication of Propositions 1 and 2 is that the WLS-based initialization will tend to outperform the OLS-based method as the length of the training sample increases.

3.3 Estimation-based methods

Another approach to the initialization of learning coefficients is to add the initials to the set of the model's parameters, and estimate them jointly. The idea can be traced back to the landmark work by Sargent et al. (2006), where the estimation of the monetary authority's initial and consecutive stream of beliefs provided evidence in favor of Sargent's (1999) hypothesis on the "Conquest of American Inflation": namely, that the rise and fall of post-WWII inflation in the US can be attributed to the evolution of the monetary authority's beliefs about the trade-off between inflation and unemployment. In spite of some early criticisms (see discussion in Carboni and Ellison, 2009), the approach of joint estimation of initials has been slowly incorporated into broader applications of adaptive learning. After being hinted as a possibility in Milani (2007, p. 2071) and Huang et al. (2009, p. 397), initial attempts have focused on the effects of the joint estimation of the initial matrix of second moments, $\mathbf{R}_0$ (e.g., Milani, 2008, 2011), and more recently on the estimation of the complete set of learning initials (as in, e.g., Slobodyan and Wouters, 2012b; Gaus and Ramamurthy, 2014).

There are different alternatives for the estimation of dynamics macroeconomic models; here, for illustrative purposes, we adopt the approach of minimizing the sum of squared residuals (SSR). Namely, the estimation-based initials are obtained according to

$$\hat{\phi}_0^{SSR} = \arg\min_{\hat{\phi}_0} SSR = \sum_{i=0}^{T-1} \hat{u}_{T-i}^2, \tag{16}$$

where T stands for the number of observations used for estimation. Under our example model's ALM, the initial estimate that solves this problem is given by

$$\hat{\phi}_0^{SSR} = \sum_{i=0}^{T-1} \mathcal{H}_i y_{T-i}, \tag{17}$$

where

$$\mathcal{H}_i = \frac{\gamma\left(\lambda^{-i}\left(1+\lambda\right) - \beta\lambda\left(\lambda^{-i} - \lambda^{i}\right)\right)}{\beta\lambda\left(\lambda^{-T} - \lambda^{T}\right)}. \tag{18}$$

The MSD associated to these estimated initials are then given by

$$\begin{aligned}\mathcal{D}_{\phi_0}^{SSR} &= E\left[\left(\phi_0 - \hat{\phi}_0^{SSR}\right)^2\right], \\ &= \frac{\gamma\left(\gamma-2\right)}{\beta^2\left(\lambda^{2T}-1\right)}\sigma_u^2. \end{aligned} \tag{19}$$

Proposition 3. *Under the model given by (2)+(6), whereas increasing the length of the estimation sample improves the accuracy of SSR-based initials, the initial estimates never converge to their corresponding true values.*

Remark 3. Notice that evaluating (19) in the limit as $T \to \infty$ one finds that the higher the learning gain, the higher the asymptotic error incurred by the SSR-based initial.

Most estimation approaches share a common idea of looking for the combination of parameter values that maximize the fit of the model, or its implications, to available macroeconomic data. Hence, the joint estimation of learning initials can have an appealing motivation for providing those estimates of the initial beliefs that are the most consistent with the data according to the chosen empirical criterion. However, as we will show in our simulation analysis in the next section, the effects of incorporating the initials into estimation routines will depend on the specification of the estimation objective.

3.4 Mixed approaches

Initialization methods can have several nuances that may not be, strictly speaking, reflected into the classes we proposed above. Particularly, there are many possibilities involving a mixture of the different approaches. For example, the REE-based initials could be computed on the basis of estimates of the model parameters obtained using data solely from the training sample. A similar approach has been used in Slobodyan and Wouters (2012b), though adding the OLS-based method to the mixture: after estimates of the model under RE are obtained, using either the training or the whole sample, the initials are set to the REE-implied OLS estimates of agents' PLM. Another example is given in Milani (2011), where the mix is between the WLS-based method and the estimation-based approach: for every draw in the Bayesian estimation routine, a training sample of observations is used to compute the initial matrix of second moments according to (12), plugging in the corresponding estimated learning gain.

Another approach, proposed by Berardi and Galimberti (2017), relies on the use of Kalman smoothing within a sample of training data in order to accelerate the convergence of the WLS-based initialization method. Although this approach requires additional computations, it has been found to provide important speed improvements under alternative specifications of the learning mechanism, such as the Stochastic Gradient (SG) algorithm (Barucci and Landi, 1997; Evans and Honkapohja, 1998), which is obtained by replacing $\mathbf{R}_t^{-1}$ by an identity matrix in the LS specification of (4). Hence, the SG does not benefit from the LS "normalization" step given

by the inverse of the matrix of second moments, which prevents the use of a diffuse prior on its initialization.

Finally, although we have focused our discussion on the use of actual data on the variables included in agents' PLM, another alternative is to use data from survey-based forecasts in order to get information about the initial conditions. Data on survey forecasts have been broadly taken as proxy for agents' actual expectations. In most of the cases, the initialization methods we discussed above can be adjusted to take advantage of this information. For example, the REE-based initials can be calculated using model's estimates obtained by replacing expectation terms by direct measures from surveys (see, e.g., Orphanides and Williams, 2005a). Learning initials consistent with surveys' information can also be obtained by adjusting the estimation-based method to maximize the fit of the forecasts implied by the learning estimates to those obtained from survey forecasts (Pfajfar and Santoro, 2010). Although we recognize the potential value of these alternatives, we restrict the scope of our analysis to the definitions of initialization methods covered by our classification.

4 Simulation Analysis

4.1 Baseline Phillips curve model

In order to shed further light on the comparative evaluation of the initialization methods reviewed above, we now analyze their quantitative properties with simulations. To provide a meaningful economic example, we focus on a standard New Keynesian Phillips curve (NKPC) model, given by

$$\pi_t = \beta \pi_{t+1}^e + \lambda x_t + u_t, \tag{20}$$

$$x_t = \rho x_{t-1} + v_t, \tag{21}$$

where π_t is inflation, π_{t+1}^e represents agents' expectations for next period inflation, x_t is a proxy for real marginal cost, and u_t is a disturbance which can be interpreted as a measurement error or as an unobserved cost-push shock. The parameters in (20) are taken as semi-structural in the sense that they can be associated to deeper structural parameters of a microfounded model (see, e.g., Mavroeidis et al., 2014). Particularly, β is the subjective discount factor and

$$\lambda = \frac{(1-\theta)(1-\theta\beta)}{\theta}\kappa, \tag{22}$$

where $\theta \in (0,1)$ represents the fraction of firms that cannot change their prices in any given period, i.e., an index of price rigidity under the Calvo framework, and $\kappa \leq 1$ is a function of the labor elasticity of production and the price elasticity of demand.

The RE solution of this model is given by

$$\pi_t = \frac{\lambda}{1-\beta\rho} x_t + u_t. \tag{23}$$

It can be shown that this equilibrium is E-stable if $\beta\rho < 1$, a condition that is automatically met under the usual assumptions that $0 < \beta < 1$ and $|\rho| < 1$ (see Evans and Honkapohja, 2001, pp. 198-200). Consistent with this solution, under adaptive learning agents form expectations according to a PLM given by

$$\pi_t = \phi_{t-1} x_t + z_t, \tag{24}$$

where ϕ_t is a parameter estimated with the univariate version of the LS algorithm given by (4)-(5), also substituting $y_t \equiv \pi_t$. Iterating (24) forward and substituting the expectations in (20) we obtain the ALM under learning

$$\pi_t = \left(\beta\rho\phi_{t-1} + \lambda\right) x_t + u_t. \tag{25}$$

4.2 Simulation and estimation approach

We generate 10,000 samples of artificial series of π_t and x_t assuming that $u_t \sim N\left(0, \sigma_u^2\right)$, $v_t \sim N\left(0, \sigma_v^2\right)$, and that $Correl\left(u_t, v_t\right) = 0$. The number of observations used for the learning initialization, in the training sample-based methods, and the estimation of the model's parameters will be a dimension of our analysis, but in general we simulate the model for 10,000 observations and assume the sample of data available for estimation starts at the 10,001 observation, i.e., $t = 0$ is observation 10,000 in our artificial series. The model parameters are set to $\beta = 0.99$, $\theta = 0.65$, $\kappa = 0.25$, $\rho = 0.9$, $\sigma_u^2 = 3$, $\sigma_v^2 = 1$, whereas for the learning gain we evaluate two options, $\gamma_1 = 0.02$ and $\gamma_2 = 0.10$[4].

In order to estimate the model and learning parameters, we follow a generalized method of moments (GMM) approach. Following Chevillon et al. (2010), we obtain the moment conditions from the common assumption that the unobserved disturbance term, u_t in our model, is a martingale difference sequence, which means $E_{t-1}\left[u_t\right] = 0$. For a given set of pre-determined instruments, the model's ALM under learning is used to derive the associated GMM objective function, which is minimized using standard numerical estimation methods[5].

One important dimension of our analysis of estimated initials is the choice of instruments used for the definition of the estimation criterion. According to Proposition 3, the use of a squared residuals criterion provides interesting results with respect to the accuracy of estimated initials as the estimation sample size grows. Nevertheless, the SSR criterion is not implicit in

[4]Our findings are qualitatively insensitive to these choices of parameters values, but not quantitatively. As evidenced in (10), though under a simpler model specification, the impact of initialization errors over the accuracy of the model's predictions is positively related to the magnitude of the parameter associated with the forward-looking term, β.

[5]Details about this estimation procedure are provided in Appendix B.

the GMM estimation approach with pre-determined instruments, and, as we will show, this can turn the estimation approach susceptible to severe initialization biases for large samples. Hence, we consider two estimation approaches. The first, which we denote as *standard*, is based on the use of four lags of the model residuals as instrument; this approach was proposed by Chevillon et al. (2010) in order to deal with the weak identification and persistence problems that are induced by adaptive learning[6]. The second approach, which we call the *augmented* method, is to use both four lags and the contemporaneous values of the model residuals as instruments, so as to enforce the SSR criterion in the estimation of the model.

4.3 Results

We conduct two sets of simulation exercises comparing the accuracy of the initialization methods, focusing on the PLM coefficient initial[7], ϕ_0, and how the different initials affect the accuracy of estimates of other model's parameters, such as the learning gain, γ, and the price rigidity parameter, θ. First, we look at the case where only the learning initial is estimated, i.e., both γ and θ are assumed to be known. This corresponds to the analysis carried out analytically in section §3 for the simpler example model, and should therefore be informative about how our previous conclusions extrapolate to more realistic models. Second, we consider the more realistic case where the model parameters need to be estimated jointly with the determination of the learning initials. Importantly, in the latter case the pre-determined initials are also conditioned on the estimates of the model parameters relevant to their determination.

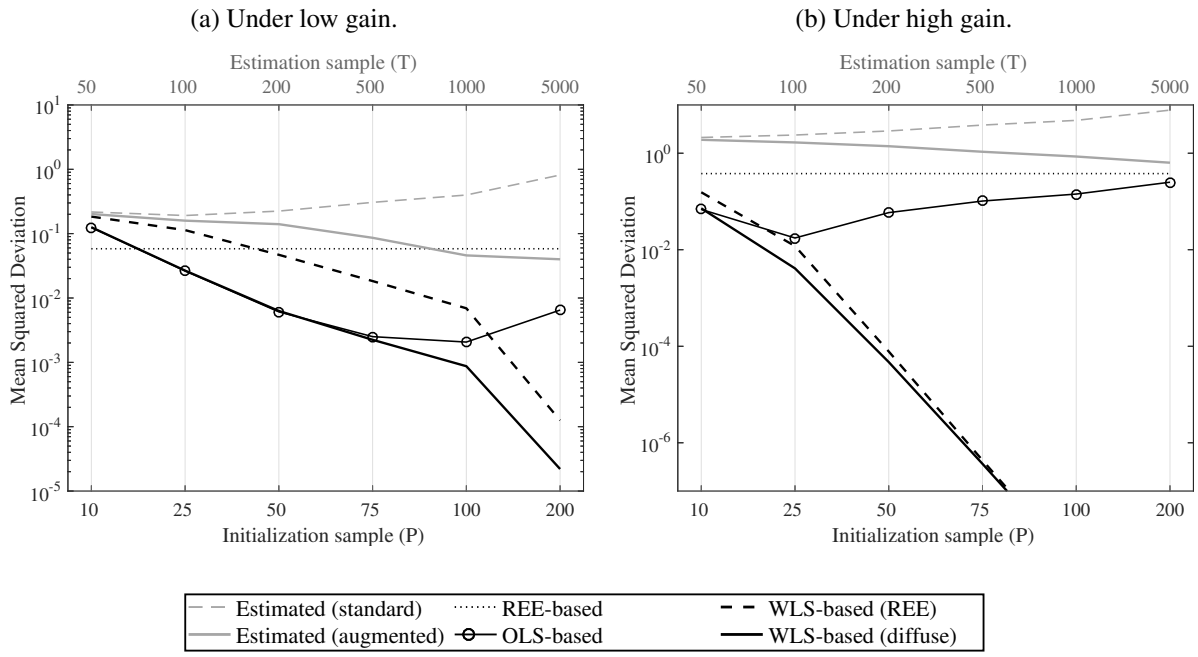
Under the REE-based method, based on the true model's parameter values, the initials are given by $\phi_0^{REE} \simeq 0.44$, and $R_0^{REE} \simeq 5.26$. For the methods based on a training sample, i.e., the OLS and the WLS initials, we first set aside an initial portion of the simulated series, $\{\pi_t, x_t\}_{t=1-P}^{0}$, and then compute the initials based on (12)-(13), varying the size of the training sample over $P = \{10, 25, 50, 75, 100, 200\}$. For the WLS case we explore two alternative specifications of the prior $R_\emptyset$, namely, a REE prior with $R_\emptyset = R_0^{REE}$, and a diffuse prior with $R_\emptyset = 0$. For the estimation-based initials we also consider two options: the standard GMM approach, using four lags of the estimated disturbances as instrument; and the augmented approach, which adds the contemporaneous disturbances to the former[8]. In order to evaluate whether the sample size has any effect over the estimates we vary it over $T = \{50, 100, 200, 500, 1000, 5000\}$.

[6]We have also experimented with the more traditional approach of taking lags of π_t and x_t as instruments, and found such alternatives to have no major effects over our main conclusions about initial estimates.

[7]Results on the accuracy of R_0, which is also dependent on the initialization method, are available upon request.

[8]The initial of the matrix of second moment, R_0, is fixed to its true value for the results reported on the estimated initials. Overall, adding R_0 to the estimation problem deteriorates the accuracy of the estimates for the PLM coefficient initial. These results are available upon request.

Figure 1: Mean squared deviations of initials with only ϕ_0 estimated.



Notes: MSDs are depicted in a logarithmic scale (vertical axis) and are based on averaged statistics over 10,000 simulations of the baseline Phillips curve model. The low and high gains are set to $\gamma_1 = 0.02$ and $\gamma_2 = 0.10$, respectively. Estimated initials are obtained by GMM using the estimation sample indicated in the upper horizontal axis, whereas the training sample-based initials use the initialization samples indicated in the lower horizontal axis.

4.3.1 Initials accuracy under known model parameters

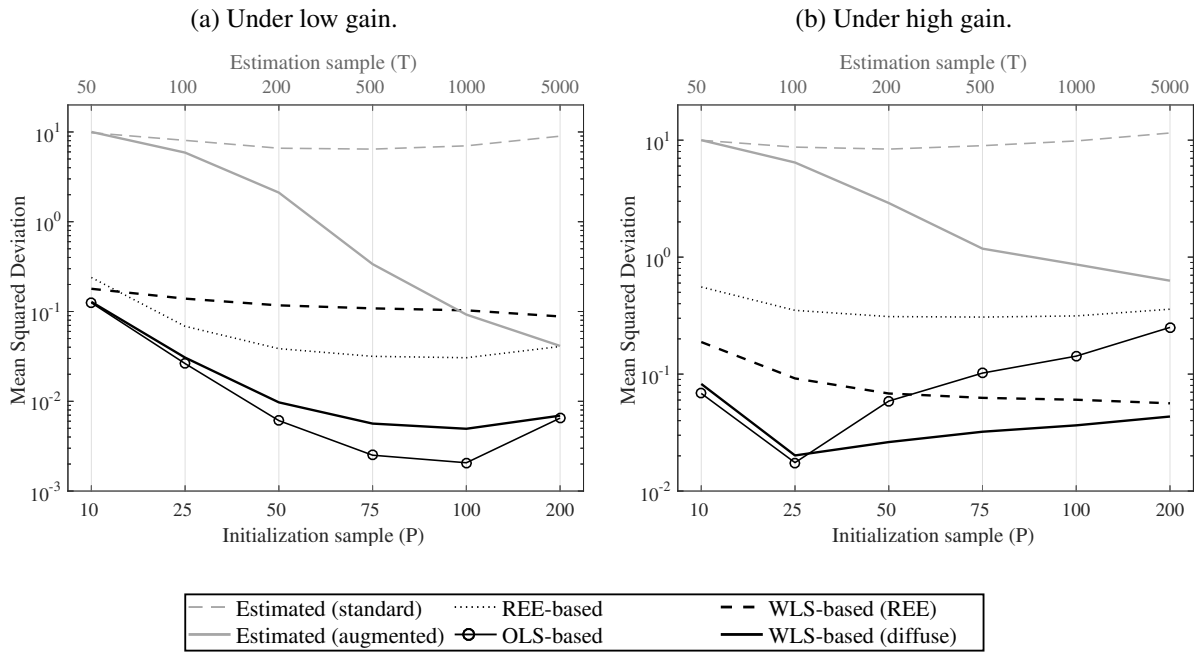
We begin looking at the case where only the learning initial is estimated. The MSDs for the initialization methods under analysis are presented in Figure 1 for the two different gain values. We draw the following observations from these results:

1. The REE-based initials are overall the less accurate among the pre-determined initials. Also, the performance of the REE-based initial deteriorates substantially for the higher gain calibration. This last result is in agreement with the result (see Eq. 8) that higher learning gains lead to noisier estimates of agents' PLM parameters, which drive out-of-equilibrium dynamics farther from the REE implied parameter values.

2. Whereas there is little difference between the OLS and the WLS initials under the lower gain calibration, the latter is clearly the best performing method under the higher gain. Also, notice that increasing the size of the training sample always improves the performance of the WLS initials, while the relationship is not monotonic for the OLS initials.

3. Between the two alternative specifications of the WLS prior on the learning coefficients uncertainty, the diffuse prior provides the most accurate initial estimates, though not in terms of the initial for the regressors' variance, where the REE prior obtains a better fit.

Figure 2: Mean squared deviations of initials with ϕ_0, γ and θ jointly estimated.



Notes: See notes of Figure 1. The estimates of γ used for the WLS-based initials are those obtained under the augmented estimation approach with $T = 200$.

Hence, the use of a diffuse prior provides an interesting way to speed up the convergence of the learning estimates within the training sample.

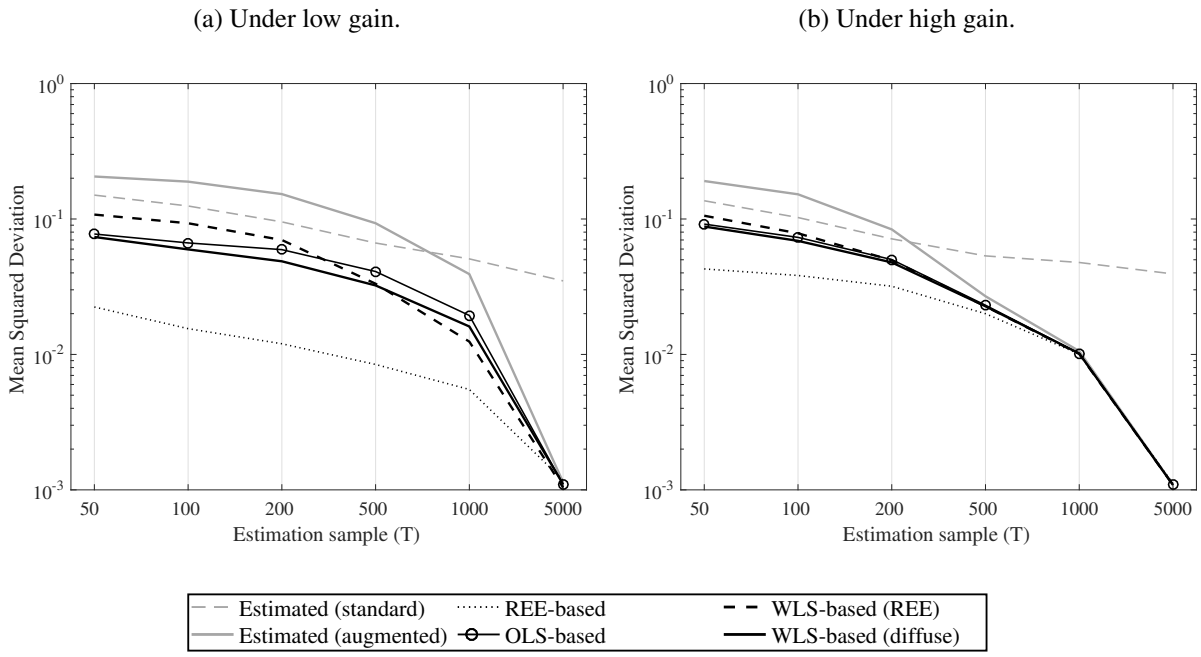
4. The results for the estimated initials are clearly dependent on the specification of the estimation criterion. Particularly, under the standard GMM approach larger estimation samples have a perverse effect on the accuracy of the estimated initials. Adding the squared residuals criterion to the estimation then brings significant improvements to the accuracy of this method, also avoiding the negative effects from the estimation sample size.

4.3.2 Initials accuracy with estimated model parameters

We now look at the more realistic case where model parameters need to be estimated together with the learning initial. Naturally, lack of knowledge about the true learning gain and the true index of price rigidity should affect the performance not only of the estimated initials, by adding more parameters to the estimation problem, but also the performance of those methods that are dependent on the model parameters, i.e., θ for the REE-based method, and γ for the WLS-based method. Figure 2 presents the results obtained under these circumstances[9] and focusing on the same initialization methods and gain calibrations evaluated above.

[9]Results for the addition of each of these parameters separately to the estimation problem are presented in Appendix C.2.

Figure 3: Mean squared deviations of estimates of price rigidity index, θ.



Notes: See notes of Figure 1. The estimates refer to the exercise with the joint estimation of γ, θ and the corresponding initial. The OLS and the WLS-based initials use a training sample of $P = 75$ observations.

The main observation sprouting from these results relates to the sensitiveness of some predetermined initials to the estimation of the model parameters. Particularly, we find that the inaccuracies in the estimation of γ can hold back the WLS convergence to the true learning coefficients within both the training sample and the estimation sample that follows the initialization. Besides, because the OLS-based method is not sensitive to any of the estimated model parameters, it yields performances identical to those observed without the estimation of other parameters. As in the previous exercise, the OLS-based initials can yield good convergence performance under small training samples, but they are not robust to variations in the underlying learning gain. With respect to the joint estimation of the initials we draw a similar conclusion to that obtained from the exercise under known model parameters: the augmented estimation approach is outperforming the standard approach.

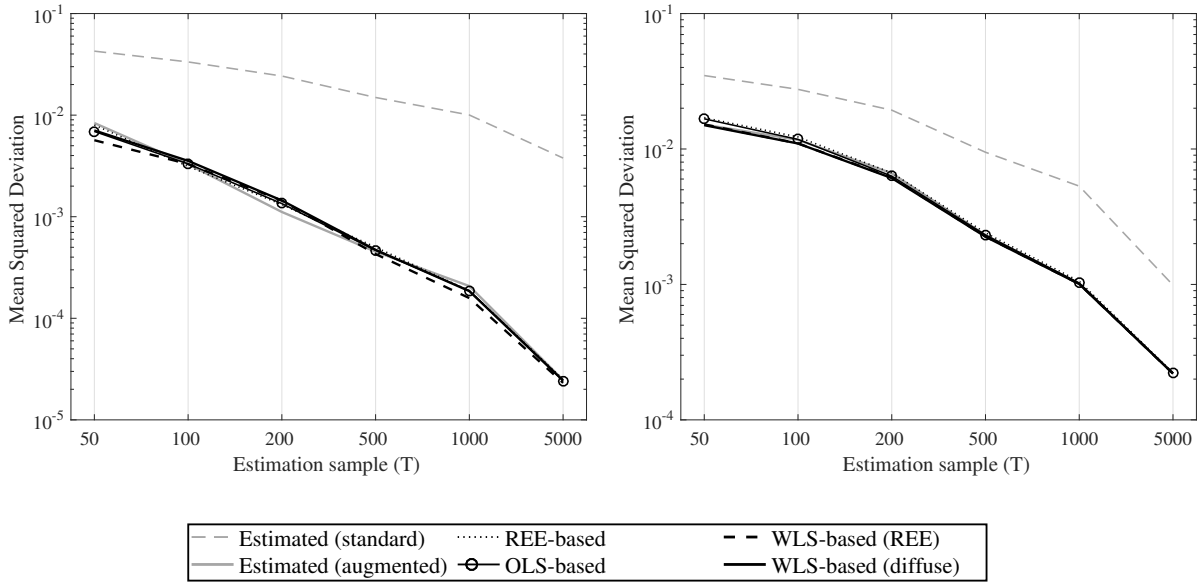
4.3.3 Initials effects over accuracy of estimated model parameters

Perhaps a more interesting question for the applied researcher is how much the errors implied by each initialization method can affect the estimates of the other model parameters. In Figures 3 and 4 we report results on the accuracy of the estimates for the parameters θ and γ, respectively, also referring to the same simulations and estimations analyzed above.

Overall we find that all estimates of the model parameters tend to converge to their corresponding true values as the estimation sample size increases, confirming the consistency property of the adopted GMM estimators. However, we also observe different rates of convergence,

Figure 4: Mean squared deviations of estimates of learning gain, γ.

(a) Under low gain. (b) Under high gain.



Notes: See notes of Figures 1 and 3.

depending on the initialization method and the estimation approach. For the estimation of θ, the REE-based initial is performing remarkably well, whereas the training sample-based methods show a slower convergence depending on the underlying gain calibration. For the estimation of γ, the pre-determined initials are performing similarly.

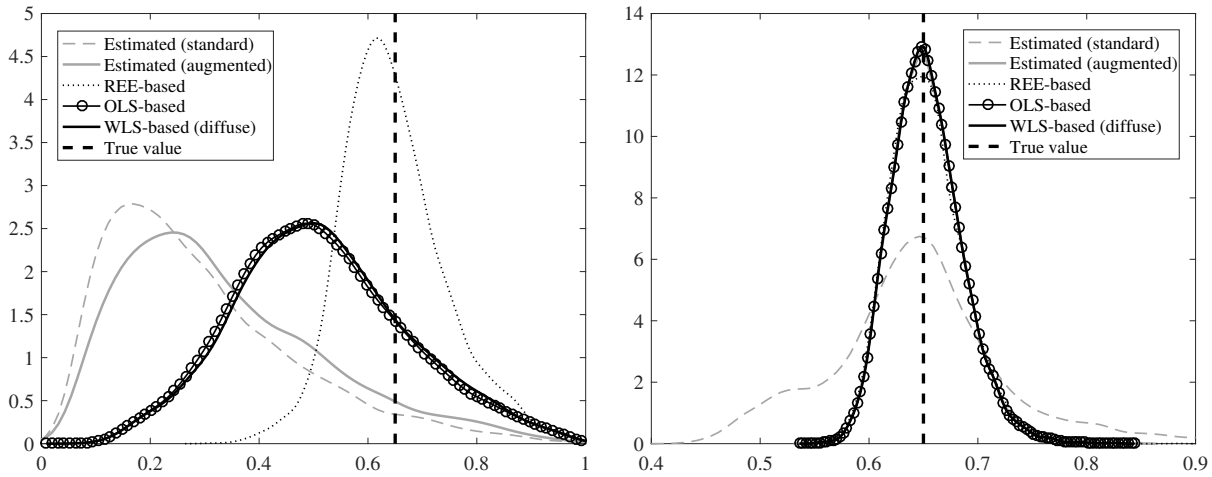
But it is with respect to the estimation approach that we observe the greatest disparities. Whereas the augmented approach is generally preferred for large samples, its performance for the estimation of θ with smaller samples is slightly poorer relative to the standard approach. Although this result cautions against the use of non-predetermined instruments in the estimation of structural parameters in models with learning, it is also clear that our augmented approach brings substantial improvements to all estimates as the sample size grows[10].

Nevertheless, the pre-determined methods are generally showing a better performance than that obtained under the joint estimation of the initials. This result is also evident in Figure 5, where we look at the densities of θ estimates when the gain is assumed to be known. Corroborating previous findings in the literature (e.g., Chevillon et al., 2010), in the left hand side (l.h.s.) panel of Figure 5 we observe that learning induces non-standard distributions in finite sample estimates of model parameters. Interestingly, we find that the joint estimation of the learning initials can lead to much stronger deviations from asymptotic distributions. Moreover, notice that even with a larger estimation sample, as in the right hand side (r.h.s.) panel of Figure 5, the θ estimates are still more dispersed under the case where the initial was jointly

[10]Another noteworthy caution against the augmented approach under small estimation samples is that numerical estimation instabilities were observed more frequently than in the standard approach for the estimation of the learning gain, as reported in Table 2 in the Appendix.

Figure 5: Densities of θ estimates over simulations.

(a) Small estimation sample (T=50). (b) Big estimation sample (T=5000).



Notes: Densities estimated using the normal kernel smoothing function over 10,000 simulations of the baseline Phillips curve model under the low gain setup, $\gamma_1 = 0.02$. Simulations with exact boundary estimates are discarded (see a report on these cases in the Appendix). The estimates refer to the exercise where only θ and ϕ_0 are estimated.

estimated using the standard approach than under the cases with pre-determined initials and the augmented estimation approach.

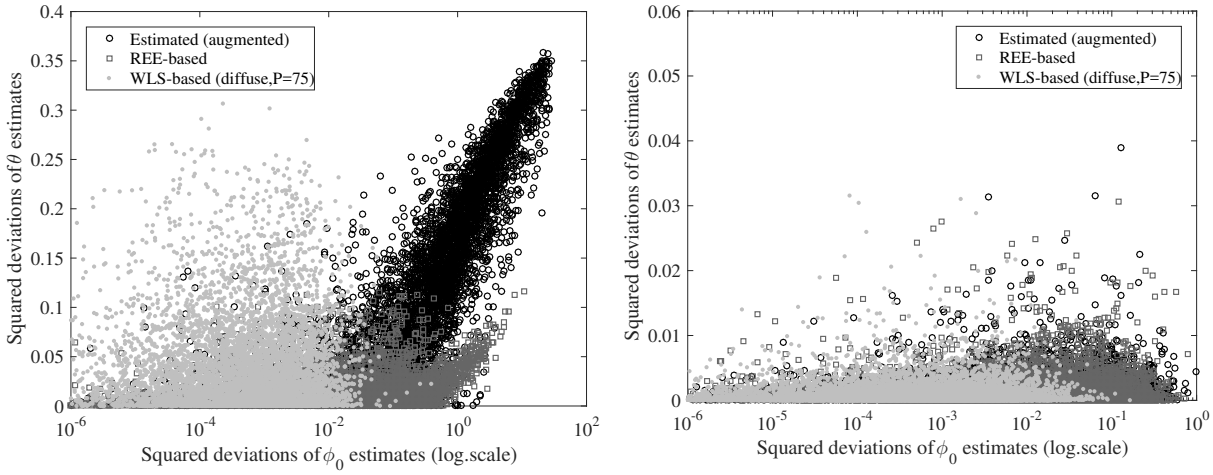
To better understand the impact of initialization errors over other model parameters estimation errors, we present in Figure 6 scatter diagrams relating the squared deviations of the initial estimates and those of θ estimates across the many simulations conducted with the joint estimation of these parameters. Importantly, we observe that the effects of the initialization errors depend on whether the initials are jointly estimated with the model parameter or pre-determined. Particularly for small estimation samples, in the l.h.s. panel of Figure 6, the joint estimation of the initial induces a positive correlation between the initials and the model parameter's estimation errors. This negative spillover effect naturally disappears as the estimation sample increases, and both the model parameter and the estimated initials get more accurate.

4.4 Summary and discussion

We draw several important conclusions regarding the performance of the different initialization methods from our simulation analysis. First, our results under the assumption of known model parameters provided robustness to our analytical findings obtained under the simpler example model of the previous section. Particularly, we find that the WLS-based initials, with diffuse priors, show the best performance in terms of initials accuracy, as long as "enough" observations are set aside for the training sample. The number of "enough" observations depends on the gain calibration, where smaller gains require longer training samples to guarantee the learning algorithm's convergence.

Figure 6: Relationship between initialization errors and model estimation errors.

(a) Small estimation sample (T=50). (b) Big estimation sample (T=5000).



Notes: The scatter diagrams are based on squared deviations between the indicated parameters estimates (ϕ_0 in the horizontal axis, and θ in the vertical one) and their corresponding true value over 10,000 simulations of the baseline Phillips curve model under the low gain setup, $\gamma_1 = 0.02$. Simulations with exact boundary estimates are discarded (see a report on these cases in the Appendix). The estimates refer to the exercise where only θ and ϕ_0 are estimated.

Second, when model parameters require estimation we found that the performance of the initialization methods is sensitive to the accuracy of the estimated parameters. However, the sensitiveness of pre-determined initials to the estimation approach was much weaker than that of the jointly estimated initials. Here, the traditional focus on pre-determined instruments was found to hinder the consistency of the initial estimates, which tended to deteriorate as the sample size increased. Our proposal to include a squared residual term in the estimation criterion found some success in reverting that relationship. One explanation for that result is that the SSR criterion penalizes more strongly the accumulation of learning estimation errors: the larger the estimation sample, the higher the precision attained by the learning estimates by the end of the sample, and with that, the more evident become the initialization errors (relative to end-of-sample errors). Without the SSR criterion, in contrast, the relative weights given to initialization errors only tended to decrease as the estimation sample size increased, reducing the relevance of initials and therefore turning their estimation indeterminate.

Notwithstanding, even under the augmented estimation approach the joint estimation of initials was found to have a negative spillover effect on the accuracy of model parameter estimates, particularly with small estimation samples. This spillover effect was not observed under the pre-determined initials. Hence, we conclude that their usage must be preferred; particularly, our evidence indicates that WLS-based method may be favored on the grounds of its greater robustness to uncertainties regarding the true values of the learning gain.

5 Empirical Application

5.1 Hybrid Phillips curve model

In order to evaluate the relevance of using different initialization methods, we now pursue an empirical application augmenting our baseline model and estimating it with US macroeconomic data. We follow Gali and Gertler (1999) and estimate a hybrid NKPC model given by

$$\pi_t = \psi_f \pi_{t+1}^e + \psi_b \pi_{t-1} + \delta x_t + \eta_t, \tag{26}$$

$$x_t = \rho x_{t-1} + \nu_t, \tag{27}$$

with

$$\psi_f = \beta\theta\zeta^{-1},\ \psi_b = \omega\zeta^{-1},\ \delta = (1-\omega)(1-\theta)(1-\beta\theta)\zeta^{-1}, \tag{28}$$

$$\zeta = \theta + \omega(1-\theta(1-\beta)), \tag{29}$$

where ω is the fraction of firms that set their prices according to a backward looking rule of thumb, and the remaining parameters have the same interpretation as in the baseline specification.

Assuming the PLM is given by

$$\pi_t = a + b\pi_{t-1} + cx_t + u_t, \tag{30}$$

and that π_t is unknown when agents form expectations at period t,

$$\pi_{t+1}^e = E_t[\pi_{t+1}] = a(1+b) + b^2\pi_{t-1} + c(b+\rho)x_t. \tag{31}$$

The REE is given by the values of a, b, and c that solve the following equalities

$$a = \psi_f a(1+b),\ b = \psi_f b^2 + \psi_b,\ c = \psi_f bc + \psi_f \rho c + \delta. \tag{32}$$

Solutions to this system are characterized in the Appendix A.5, where we also show that there is an unique E-stable REE in this model, given by $a^* = 0$, $b^* = \left(1 - \sqrt{1 - 4\psi_f\psi_b}\right)/2\psi_f$, and $c^* = \delta/(1 - \psi_f(b^* + \rho))$. Under adaptive learning, agents form expectations using estimates of these parameters obtained according to the LS algorithm of (4)-(5), where $\hat{\boldsymbol{\phi}}_t \equiv \left(\hat{a}_t, \hat{b}_t, \hat{c}_t\right)'$, $\mathbf{x}_t \equiv (1, \pi_{t-1}, x_t)'$ and $y_t \equiv \pi_t$.

5.2 Data and estimation approach

We use quarterly US data covering the period from 1947q1 to 2016q3. To measure inflation we use CPI inflation, whereas for the forcing variable, x_t, we use non-farm business sector labor

shares. To remove trends in the latter we obtain gap measures using the Hodrick-Prescott (HP) filter. For simplicity, we are neglecting real-time data issues by focusing on a unique snapshot of the realization of these series. All our data series are obtained from the FRED database of the St. Louis Fed.

We adopt a GMM estimation approach similar to that used in our simulation analysis. Here, the set of parameters determining inflation is given by $\Theta = (\beta, \theta, \omega, \rho, \gamma, \mathbf{\Phi}_0, \mathbf{R}_0)$. To facilitate estimation we fix some of these parameters to values we consider reasonable: $\beta = 0.99$, $\rho = 0.60$, the latter corresponding (approximately) to the OLS estimate of (27). We experiment with both estimation approaches discussed in our simulation exercise. Under the standard estimation approach the instruments are given by a constant plus four lags of the estimation residuals, the HP-filtered output gap on real GDP, and a long-short interest rate spread given by the difference between the 10-year Treasury Bill rate and the Federal Funds rate. Under the augmented approach we then add the contemporaneous estimation errors to the former list of instruments[11].

In order to conduct a comparative evaluation of the estimates associated to the different initialization methods, we fix the estimation sample to start from 1972q2. That leaves up to 100 observations for the training sample-based initializations. For the cases of the REE-based and the WLS-based methods, the initials are re-calculated within the estimation routine in order to take into account the estimates of the model parameters and the learning gain, respectively. We use a diffuse prior for the WLS-based method.

We also benchmark our estimates of the model under learning with corresponding estimates under RE. For that purpose we follow the approach of replacing the expectation term in (26) by actual observations of next period inflation. Imposing the same identifying assumption we used under learning, which in the present context is given by $E_{t-1}[\eta_t] = 0$, leads to moment conditions of the form of (52), except that the residual term now also includes an one-step-ahead inflation forecast error (see Mavroeidis et al., 2014, pp. 133-4). Finally, as instruments for the RE estimation we use four lags of inflation and the labor income share gap instead of the lagged residuals.

5.3 Robust stable learning estimates

Models with adaptive learning may generate unrealistic dynamics due to instabilities in the recursive estimation of the learning coefficients. That is particularly relevant for the case of constant gain learning: the higher the value of the learning gain, the higher the chances that a sequence of abnormal forecast errors will push agents' PLM estimates outside the region

[11] In a previous version of this paper (Berardi and Galimberti, 2016) we also included an unrestricted constant in the model specification, which turned out to result in unreasonable and less stable parameter estimates under learning. One possible explanation for this issue relates to the small sample sensitiveness of the GMM estimator to the normalization of orthogonality conditions with nonlinearities (see Fuhrer et al., 1995). Hence, in the current version we decided to include the unrestricted constant only for the estimation of the model under RE.

of stability in the parameters' space. Therefore, the estimation of models under learning can be extremely sensitive to the sample of observations and the range of values allowed for the learning gain.

To deal with these potential sources of instabilities we develop a robust stability estimation approach. Namely, we impose additional constraints to the estimates of the learning gains so as to ensure the stability of learning[12]. Importantly, due to self-referentiality, these constraints can depend on the values of the model parameters. Considering that we are also estimating θ and ω we draw the upper bounds to be imposed on the estimation of the learning gain through simulations of the model for different combinations of those parameters[13]. The resulting upper bounds are presented in Figure 7. We observe that the model becomes more sensitive to the learning gain as the model parameters get closer to their boundary values. To understand this result notice that from (26): (i) as $\theta, \omega \to 1$, inflation converges to a random walk process because $\delta \to 0$ and $\psi_f + \psi_b \to 1$; and, (ii) as $\theta, \omega \to 0$, the model becomes indeterminate because $\zeta \to 0$.

Another mechanism that is often coupled to the learning algorithm in order to prevent the learning estimates to diverge is the so-called "projection facility". Whenever the learning estimates leave a bounded region in the parameters space, the projection facility is activated in order to contain the escape. This device is particularly useful for the cases of PLMs containing lagged endogenous variables, where the definition of parameters' bounds can be easily derived from stationarity conditions. Here, for the estimation of the hybrid NKPC we also adopt a projection facility that is activated whenever $\left|\hat{b}_t\right| \geq 1.5$, in which case the learning algorithm does not update the learning estimates and discards the associated observation[14].

5.4 Results

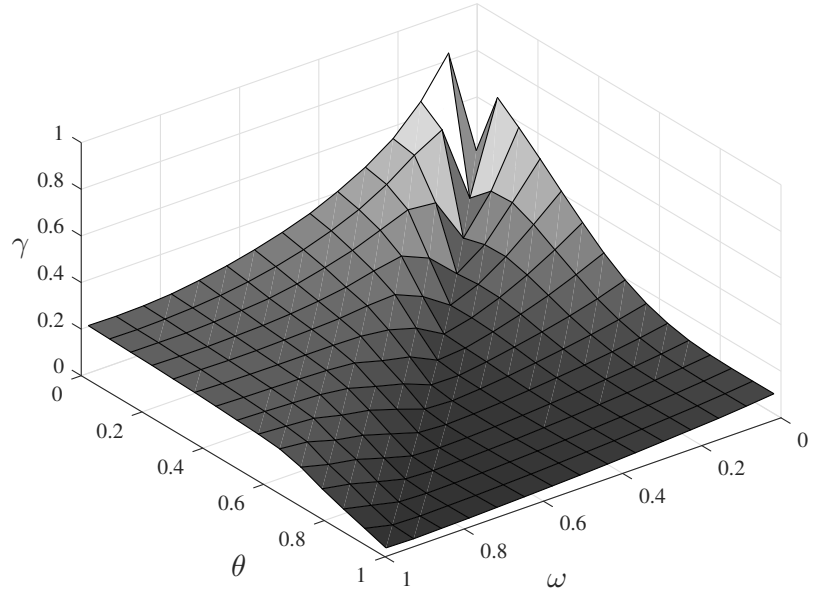
The results obtained with the GMM estimation of the hybrid NKPC model are presented in Table 1. The RE benchmark estimates seem to be in accordance to previous estimates in the literature (e.g., Gali and Gertler, 1999) where the Calvo's index of price stickiness, θ, is found to be greater than the fraction of backward looking price setters, ω. This result is also robust to the introduction of learning, although with some variation in the point estimates across the estimation and initialization approaches. Particularly, we find that the degree of price stickiness

[12] Although we are not aware of any previous attempt to introduce such robustness conditions in the estimation of learning models in the literature, this approach is inspired by Evans and Honkapohja (2009), who showed that constant gain learning can turn an otherwise learnable equilibrium unstable, depending on the value of the gain. We thank an Associate Editor for pointing us towards this direction.

[13] See the Appendix C.3 for details about these simulations.

[14] Although stationarity of (30) requires that $|b| < 1$, we adopt a critical value of 1.5 so as to allow the algorithm some space to return to the stability region on its own.

Figure 7: Learning gain simulated upper bounds for hybrid New Keynesian Phillips curve model.



Notes: Upper bounds obtained from 1,000 simulations of model (26) under learning over a grid of 15 values of $\theta = \{0.01, 0.08, ..., 0.91, 0.99\}$ and 15 values of $\omega = \{0.01, 0.08, ..., 0.91, 0.99\}$. The reported upper bounds refer to averaged maximum gain values for which the simulated learning estimates of $\hat{b}_t$ remained between ± 1.5 over 1,000 periods. See the Appendix C.3 for further details.

tended to decrease with the introduction of learning[15], which is consistent with the findings of Milani (2007).

We also find that the adaptive learning approach provides a superior fit to the data. That is clearly evidenced by looking at the inflation expectations implied by the different model estimates, depicted in Figure 8. These expectations are obtained from the assumed PLM, as in (31), taking the learning initials as point of departure and updating these estimates according to the learning algorithm and the corresponding data. As expected, the constant-gain learning mechanism does a better job than the RE approach in tracking the fluctuations of inflation over the sample of data.

Comparing the differences due to changes in the estimation and the initialization methods we find that the augmented estimation approach provides more reliable estimates. Particularly, we observe that there is much less variation on the estimates across the initializations when the parameters are estimated under the augmented rather than the standard approach. E.g., ω varies between 0.01 (WLS- and OLS-based initial) and 0.47 (estimated initial) under the standard estimation approach, while under the augmented approach ω is narrower identified between 0.31 (estimated initial) and 0.44 (WLS-based initial); these variations have important implications for the impact of the forcing variable on inflation, as evidenced by the implied δ estimates in the last column of Table 1, which is a crucial parameter for policymakers. The same applies to the other estimated parameters, except for the learning initials, where the estimates of a_0, b_0, and c_0 still show some variation across the initialization methods.

Regarding the initials, empirically, it is difficult to assess which of the initialization methods provided the most accurate initial estimates. In our simulation analysis we have argued in favor of the WLS-based estimates for providing the estimates that are most coherent to the ongoing learning process at the beginning of the estimation sample. That seems to be the case here, as the OLS-based initials point to similar estimates.

In order to further illustrate the differences in the implied expectations across the initialization methods, we re-estimate the model under varying sample beginnings and compare their correponding expectations in Figure 9. Particularly, we focus on augmented estimates obtained with samples beginning around the Great Inflation episode in the US: (i) from 1966q4, capturing the preceding period; (ii) from 1972q2, taken as the onset of the referred episode; and, (iii) 1985q1, leaving behind this period of greater volatility and focusing on the Great Moderation period. Here we can clearly see the potential biases that the joint estimation of the learning initials can introduce in the modelled expectations: estimating the initials with data starting from 1972q2 leads to an expected inflation of more than 3% (quarterly rate), whereas the expectations implied by an ongoing process would be around 1% for that period. The expectations

[15]The only exceptions to this finding occur under the standard estimation approach with jointly estimated and REE-based initials, but notice these estimates were not statistically significant. This finding is also robust to alternative estimation samples, as reported in Appendix C.4, although with substantial variation in the point estimates; this is not surprising considering that empirical estimates of the NKPC have been historically sensitive to data and estimation techniques (see Mavroeidis et al., 2014).

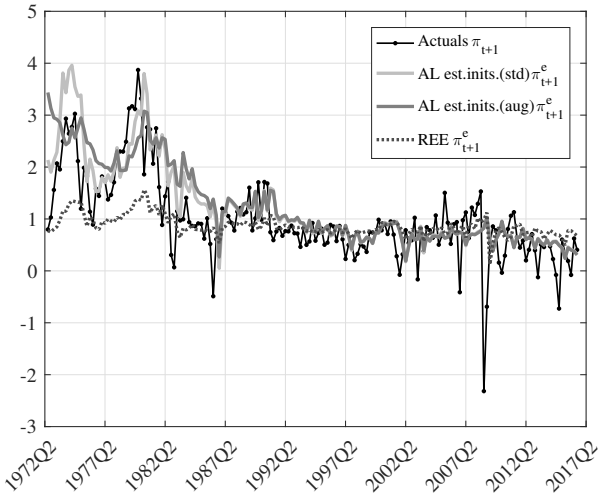
Table 1: Empirical estimates of US NKPC - 1972q2 - 2016q3.

Exercise	Parameter estimates: Structural		Learning				Implied reduced form parameters		
	θ	ω	γ	a_0	b_0	c_0	ψ_f	ψ_b	δ
Rational expectations									
	0.85	**0.40**					0.68	0.32	0.01
	(0.09)	**(0.06)**							
Learning with jointly estimated initials									
- Standard	0.93	0.47	0.04	0.79	**0.99**	0.13	0.66	0.34	0.00
	(1.07)	(0.71)	(0.03)	(0.54)	**(0.32)**	(0.26)			
- Augmented	**0.62**	**0.31**	**0.05**	2.31	0.29	-0.37	0.66	0.33	0.11
	(0.07)	**(0.11)**	**(0.02)**	(2.02)	(0.89)	(0.53)			
Learning with REE-based initials									
- Standard	0.95	0.17	**0.08**	0.00	0.18	0.01	0.84	0.15	0.00
	(0.58)	(0.10)	**(0.02)**						
- Augmented	**0.62**	**0.39**	**0.03**	0.00	0.63	0.34	0.61	0.39	0.09
	(0.11)	**(0.06)**	**(0.02)**						
Under learning with WLS-based initials									
- Standard	**0.74**	0.01	**0.11**	0.53	0.45	0.12	0.98	0.01	0.09
	(0.06)	(0.09)	**(0.02)**						
- Augmented	**0.56**	**0.44**	**0.02**	0.17	0.76	0.04	0.56	0.44	0.11
	(0.06)	**(0.05)**	**(0.01)**						
Under learning with OLS-based initials									
- Standard	**0.75**	0.01	**0.11**	0.17	0.76	0.04	0.98	0.01	0.08
	(0.07)	(0.09)	**(0.02)**						
- Augmented	**0.58**	**0.43**	**0.02**	0.17	0.76	0.04	0.57	0.43	0.10
	(0.06)	**(0.06)**	**(0.01)**						

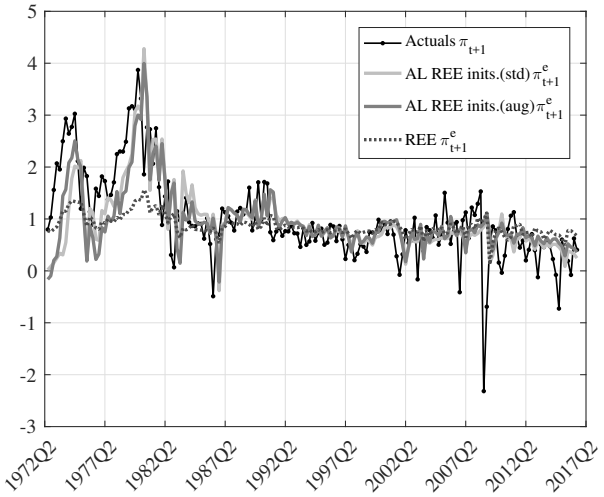
Notes: Parameters estimated by GMM, as explained in the text, using US data from 1972q2 to 2016q3. The implied reduced form parameters are obtained according to (28)-(29). For the cases with pre-determined initials, the initial learning coefficients are obtained from training samples of 75 observations over the period from 1953q3 to 1972q1. Values in parentheses are standard errors of the estimates and are computed based on numerical approximations of the objective function first derivatives. The standard errors under learning should be interpreted with caution since the estimators distribution, and corresponding test statistics, can become non-standard (see Chevillon et al., 2010). Estimates in bold are those found to be statistically different than zero at a 5% level of significance.

Figure 8: Expectations implied by model estimates from 1972q2 to 2016q3.

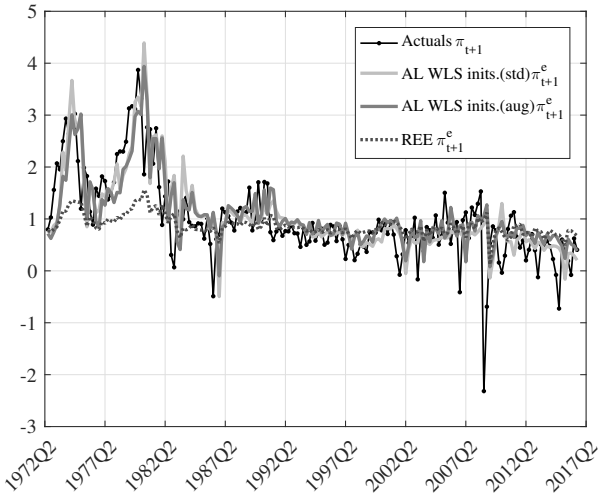
(a) Estimated initials.



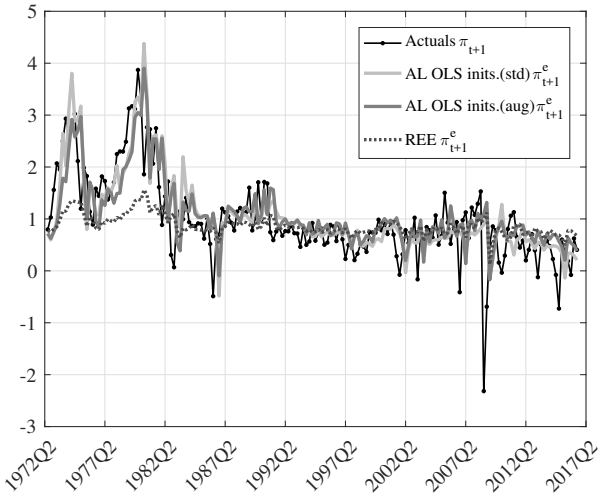
(b) REE-based initials.



(c) WLS-based initials.



(d) OLS-based initials.



Notes: The inflation expectations are computed according to (31) departing from the estimates of the learning initials and the gain reported in Table 1. For the REE expectations we use the equilibrium values implied by the structural estimates, according to (32).

implied by the REE-based initials also present some deviation, which is consistent with our previous finding that the REE fails to capture the dynamics of inflation over this period.

Summing up, we interpret these empirical results as providing further evidence in favor of the use of training samples to initialize learning estimates together with the use of the augmented estimation approach in order to estimate models with learning. The inclusion of the SSR criterion seems to play a major role in facilitating the identification of the model parameters, particularly for turning these estimates less sensitive to the learning initials. Irrespective of the initialization method, all the augmented estimates of the hybrid NKPC tell a similar story: when agents need to learn how to form their inflation expectations, our inference points to a lower degree of price stickiness than suggested by estimates under the RE assumption. Furthermore, the augmented estimates do not reject the hybrid form of the NKPC, although the effect of learning on the fraction of backward looking firms seems to depend on the initialization adopted.

6 Concluding remarks

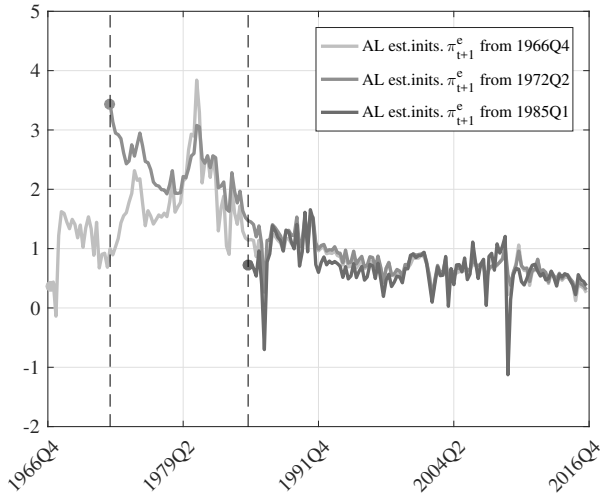
In this paper we provided a critical review on several methods previously proposed in the literature of learning and expectations in macroeconomics in order to initialize its learning algorithms. We proposed a taxonomy of initialization methods that can be broadly defined in three major classes: equilibrium-related methods, training sample-based methods, and estimation-based methods. We conducted extensive simulation exercises comparing different initialization methods that can be conceived within this classification.

Our analysis led us to draw the following recommendations. First, though equilibrium-related initialization methods seem to provide rather conservative initials, they are often incoherent with the dynamics implied by learning, particularly under high learning gains. Second, among the training sample-based methods, the use of standard OLS estimates can also turn out to provide incoherent estimates since it does not take into account the particular specification of the learning gain. Direct application of the learning algorithm into the training sample, the WLS-based method in our terminology, was overall favored by our evaluation criteria. Particularly, we found that a diffuse specification of this method leads to an accelerated convergence, facilitating the feasibility of the method in macroeconomic contexts.

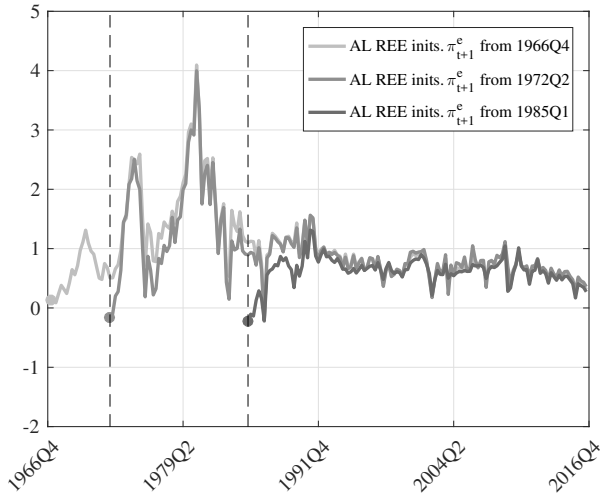
We also uncovered some problems with the approach of joint estimation of the learning initials with other model parameters. Strikingly, the accuracy of these estimates tended to deteriorate with the sample size, and, perhaps more importantly, we found evidence of spillover effects from the biases introduced by estimation of the initials into the estimates of the model's structural parameters. We proposed a solution to the first problem by including a squared residual criterion to the estimation objective; whereas this approach restored consistency to the joint estimation of the learning initials, attenuating the negative spillover effect asymptotically, the estimation of initials was still problematic under small samples.

Figure 9: Expectations implied by augmented model estimates varying the sample beginnings.

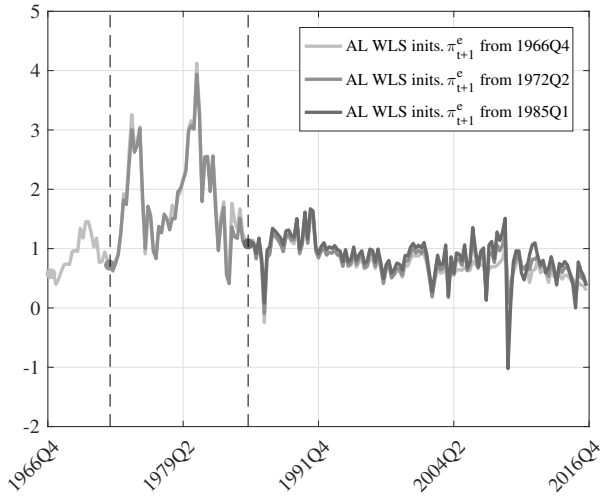
(a) Estimated initials.



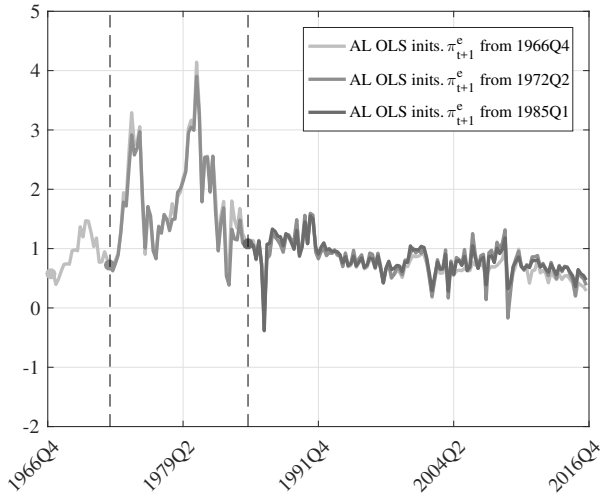
(b) REE-based initials.



(c) WLS-based initials.



(d) OLS-based initials.



Notes: The inflation expectations are computed as described in the notes of Figure 8, although based on parameter estimates obtained under the different sample beginnings as indicated in the legends. The estimates for these alternative samples are reported in Appendix C.4.

Finally, the sensitiveness of model estimates to the learning initialization method and the estimation approach was illustrated with an empirical application on the determination of US inflation rates under the New Keynesian Phillips curve framework. Our estimation approach was found to be the most robust to alternative specifications of the initials, with results indicating that the introduction of learning leads to a decrease in the estimated degree of price stickiness relative to what would otherwise be inferred under rational expectations. This finding is consistent with the idea that macroeconomic persistency is not only due to firms inability to adjust prices instantaneously, but also to the sluggish adjustment of their expectations in real-time.

A Derivations

A.1 MSPE implied by initialization error

From (2) we have that $y_{t+1} = \beta\phi_t + u_{t+1}$ and $\hat{y}_{t+1} = \beta\hat{\phi}_t$, so that the prediction error is given by

$$\begin{aligned} \hat{\Delta}_{t+1} &= y_{t+1} - \hat{y}_{t+1}, \\ &= \beta\left(\phi_t - \hat{\phi}_t\right) + u_{t+1}. \end{aligned} \tag{33}$$

Defining $\hat{\epsilon}_t = \phi_t - \hat{\phi}_t$, from (9) we find that

$$\begin{aligned} \hat{\epsilon}_t &= \phi_{t-1} + \gamma\left(y_t - \phi_{t-1}\right) - \hat{\phi}_{t-1} - \gamma\left(y_t - \hat{\phi}_{t-1}\right), \\ &= (1-\gamma)\left(\phi_{t-1} - \hat{\phi}_{t-1}\right), \\ &= (1-\gamma)\,\hat{\epsilon}_{t-1}, \end{aligned}$$

which can be solved recursively to result in $\hat{\epsilon}_t = (1-\gamma)^t\,\hat{\epsilon}_0$. Substituting this back into (33) and taking the expectation of the squared value results in

$$\begin{aligned} \hat{\Delta}_{t+1} &= \beta\,(1-\gamma)^t\,\hat{\epsilon}_0 + u_{t+1}, \\ E\left[\hat{\Delta}^2_{t+1}\right] &= \beta^2\,(1-\gamma)^{2t}\,\hat{\epsilon}_0^2 + \sigma_u^2. \end{aligned}$$

For the case where β is unknown, assuming $E\left[\hat{\beta}u_t\right] = 0$, the MSPE is then given by

$$\begin{aligned} E\left[\hat{\hat{\Delta}}^2_{t+1}\right] &= E\left[\left(\beta\phi_t - \hat{\beta}\hat{\phi}_t + u_{t+1}\right)^2\right], \\ &= E\left[\hat{\hat{\epsilon}}^2_t\right] + \sigma_u^2, \end{aligned} \tag{34}$$

where $\hat{\hat{\epsilon}}_t = \beta\phi_t - \hat{\beta}\hat{\phi}_t$, leading to

$$\hat{\hat{\epsilon}}_t = (1-\gamma)\,\hat{\hat{\epsilon}}_{t-1} + \gamma\left(\beta - \hat{\beta}\right) y_t.$$

Solving this recursively yields

$$\hat{\hat{\epsilon}}_t = (1-\gamma)^t\,\hat{\hat{\epsilon}}_0 + \gamma\left(\beta - \hat{\beta}\right)\sum_{i=0}^{t-1}(1-\gamma)^i\, y_{t-i}.$$

A.2 Long run variance of learning estimates

Substituting (2) into (4) we obtain

$$\phi_t = \delta\phi_{t-1} + \gamma u_t,$$

where $\delta = 1 - \gamma\,(1-\beta)$. This recursion is equivalent to

$$\phi_t = \delta^t\phi_0 + \gamma\sum_{i=0}^{t-1}\delta^i u_{t-i}, \tag{35}$$

In the limit, as $t \to \infty$, $E\left[\phi_\infty\right] = 0$ as long as $|\delta| < 1$. Hence, the long run variance of ϕ_t, denoted by $\overline{\sigma}^2_\phi$, is given by

$$\begin{aligned}
\overline{\sigma}^2_\phi &= \lim_{t\to\infty} E\left[\phi_t^2\right], \\
&= \lim_{t\to\infty} E\left[\delta^{2t}\phi_0^2 + 2\delta^t\phi_0\gamma\sum_{i=0}^{t-1}\delta^i u_{t-i} + \left(\gamma\sum_{i=0}^{t-1}\delta^i u_{t-i}\right)^2\right],
\end{aligned}$$

which, because u_t is assumed to be serially independent, simplifies to

$$\begin{aligned}
\overline{\sigma}^2_\phi &= \lim_{t\to\infty}\delta^{2t}\phi_0^2 + \gamma^2\sum_{i=0}^{t-1}\delta^{2i}\sigma_u^2, \\
&= \lim_{t\to\infty}\delta^{2t}\phi_0^2 + \gamma^2\sigma_u^2\frac{1-\delta^{2t}}{1-\delta^2}, \\
&= \frac{\gamma^2\sigma_u^2}{1-\delta^2}, \\
&= \frac{\gamma\sigma_u^2}{(1-\beta)\,(1+\delta)}
\end{aligned}$$

where the limit is solved under the assumption that $|\delta| < 1$. Notice that $\partial\overline{\sigma}^2_\phi/\partial\gamma > 0$, i.e., as the gain increases the dispersion of the learning estimates tends to increase as well.

A.3 MSDs of training sample-based initials

For the training sample-based initials we consider a sample of P observations, and in this case (7) translates to

$$\phi_0 = \delta^P \phi_{-P} + \gamma \sum_{i=0}^{P-1} \delta^i u_{-i}. \tag{36}$$

A.3.1 WLS-based initials

The WLS-based initial is equivalent to the application of the constant gain LS algorithm to the training sample, departing from $\hat{\phi}_{-P} = 0$ (so-called diffuse prior[16]). Yet, because the actuals are determined by the true learning estimates, (36) can not be directly employed to obtain the WLS estimates. Taking that into account we find that the WLS-based initial is given by

$$\begin{aligned}\hat{\phi}_0^{WLS} &= \gamma\beta\lambda^{P-1}\phi_{-P} + \gamma\beta\lambda^{-1}\sum_{i=1}^{P-1}\lambda^i\phi_{-i} + \gamma\sum_{i=0}^{P-1}\lambda^i u_{-i}, \\ &= \left(\delta^P - \lambda^P\right)\phi_{-P} + \gamma\sum_{i=0}^{P-1}\delta^i u_{-i},\end{aligned} \tag{37}$$

where $\lambda = 1 - \gamma$. The corresponding MSD is then given by

$$\begin{aligned}\mathcal{D}_{\phi_0}^{WLS} &= E\left[\left(\lambda^P\phi_{-P}\right)^2\right], \\ &= \lambda^{2P}E\left[\phi_t^2\right].\end{aligned} \tag{38}$$

A.3.2 OLS-based initials

The OLS-based initial is obtained by application of the Ordinary Least Squares estimator to the training sample of observations, which is equivalent to the use of (6) with a decreasing gain, $1/t$, instead of the constant, γ, and also departing from the diffuse prior, $\hat{\phi}_{-P} = 0$. The OLS-based initial is then given by

$$\begin{aligned}\hat{\phi}_0^{OLS} &= \frac{\beta}{P}\phi_{-P} + \frac{\beta}{P}\sum_{i=1}^{P-1}\phi_{-i} + \frac{1}{P}\sum_{i=0}^{P-1}u_{-i}, \\ &= \frac{\beta\left(\delta^P - 1\right)}{P\left(\delta - 1\right)}\phi_{-P} + \frac{1}{P\left(1-\beta\right)}\sum_{i=1}^{P-1}\left(1 - \beta\delta^i\right)u_{-i} + \frac{1}{P}u_0,\end{aligned} \tag{39}$$

from which the corresponding MSD follows as

$$\mathcal{D}_{\phi_0}^{OLS} = \left(\delta^P - \frac{\beta\left(\delta^P - 1\right)}{P\left(\delta - 1\right)}\right)^2 E\left[\phi_t^2\right] + \sigma_u^2\mathcal{K}, \tag{40}$$

[16] These derivations can be easily extended to the case with any other arbitrary prior, such as the REE-based prior

where

$$\mathcal{K} = \left(\gamma - \frac{1}{P}\right)^2 + \sum_{i=1}^{P-1}\left(\gamma\delta^i + \frac{\beta\delta^i - 1}{P\left(1-\beta\right)}\right)^2. \tag{41}$$

A.4 SSR estimation-based initials and associated MSD

The SSR estimation-based initials are obtained according to

$$\hat{\phi}_0^{SSR} = \arg\min_{\hat{\phi}_0} SSR = \sum_{i=0}^{T-1}\hat{u}_{T-i}^2. \tag{42}$$

Using (2) we find that the FOC to this minimization problem is given by

$$\beta\sum_{i=0}^{T-1}\frac{\partial\hat{\phi}_{T-i-1}}{\partial\hat{\phi}_0}\hat{\phi}_{T-i-1} = \sum_{i=0}^{T-1}\frac{\partial\hat{\phi}_{T-i-1}}{\partial\hat{\phi}_0}y_{T-i}. \tag{43}$$

Using (6) we find that, departing from an estimated initial and using the data available in the estimation sample,

$$\hat{\phi}_{T-i-1} = \lambda^{T-i-1}\hat{\phi}_0 + \gamma\sum_{j=0}^{T-i-2}\lambda^j y_{T-i-j-1}, \tag{44}$$

so that

$$\frac{\partial\hat{\phi}_{T-i-1}}{\partial\hat{\phi}_0} = \lambda^{T-i-1}, \tag{45}$$

where $\lambda = 1-\gamma$. Substituting (44) and (45) into (43), we find that the initial estimate is given by

$$\hat{\phi}_0^{SSR} = \sum_{i=0}^{T-1}\mathcal{H}_i y_{T-i}, \tag{46}$$

where

$$\mathcal{H}_i = \frac{\gamma\left(\lambda^{-i}\left(1+\lambda\right) - \beta\lambda\left(\lambda^{-i} - \lambda^i\right)\right)}{\beta\lambda\left(\lambda^{-T} - \lambda^T\right)}. \tag{47}$$

The MSD associated to the estimated initials can be derived by using the ALM to obtain

$$y_{T-i} = \beta\delta^{T-i-1}\phi_0 + \beta\gamma\sum_{j=0}^{T-i-2}\delta^j u_{T-i-j-1} + u_{T-i}, \tag{48}$$

which introduced into (46) leads to

$$\begin{aligned}\hat{\phi}_0^{SSR} &= \phi_0\beta\delta^{T-1}\sum_{i=0}^{T-1}\mathcal{H}_i\delta^{-i} + \beta\gamma\delta^{-1}\sum_{i=1}^{T-1}u_{T-i}\delta^i\sum_{j=0}^{i-1}\mathcal{H}_j\delta^{-j} + \sum_{i=0}^{T-1}\mathcal{H}_i u_{T-i},\\ &= \phi_0 + \frac{\gamma\left(\gamma-2\right)}{\beta\lambda\left(\lambda^T - \lambda^{-T}\right)}\sum_{i=0}^{T-1}\lambda^{-i}u_{T-i}.\end{aligned} \tag{49}$$

This expression can be used to evaluate the MSD according to

$$\begin{aligned}\mathcal{D}_{\phi}^{SSR} &= E\left[\left(\phi_0 - \hat{\phi}_0^{SSR}\right)^2\right], \\ &= E\left[\phi_0^2\right] - 2E\left[\phi_0\hat{\phi}_0\right] + E\left[\hat{\phi}_0^2\right], \\ &= E\left[\hat{\phi}_0^2\right] - E\left[\phi_0^2\right], \\ &= \frac{\gamma\left(\gamma-2\right)}{\beta^2\left(\lambda^{2T}-1\right)}\sigma_u^2. \end{aligned} \tag{50}$$

A.5 Hybrid Phillips curve REEs

There are six solutions to the RE conditions in (32). Starting with b, there are two possible solutions given by $b^{\pm} = \frac{1\pm\sqrt{1-4\psi_f\psi_b}}{2\psi_f}$. For a there are three possibilities: $a = 0$ or a is indeterminate with $\beta = 1$ and $\omega \neq 1$, or $\omega = 1$ and $\beta \neq 1$. Finally, c is uniquely determined by b. Putting these combinations together we have the following RE solutions:

$$RE.1 = \left\{a = 0,\, b^{+},\, c^{+}\right\};\; RE.2 = \left\{a = 0,\, b^{-},\, c^{-}\right\};$$

$$RE.3 = \left\{a = any,\, b^{+},\, c^{+},\, \beta = 1,\, \omega \neq 1\right\};\; RE.4 = \left\{a = any,\, b^{+},\, c^{+},\, \omega = 1,\, \beta \neq 1\right\};$$

$$RE.5 = \left\{a = any,\, b^{-},\, c^{-},\, \beta = 1,\, \omega \neq 1\right\};\; RE.6 = \left\{a = any,\, b^{-},\, c^{-},\, \omega = 1,\, \beta \neq 1\right\}.$$

Naturally, the solutions with indeterminate a can be ruled out on economic grounds: the subjective discount factor is generally assumed to be smaller than unity, whereas $\omega = 1$ would break the relationship between π_t and x_t underlying the Phillips curve.

To check for E-stability of these solutions we first define the T-mapping associated to this model:

$$T\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{bmatrix} \psi_f a\left(1+b\right) \\ \psi_f b^2 + \psi_b \\ \psi_f bc + \psi_f c\rho + \delta \end{bmatrix}.$$

E-stability requires that the eigenvalues of the Jacobian matrix of T, evaluated at the given RE solution, are smaller than unity. These eigenvalues depend only on the value of b and are given by

$$\left\{\left(1+b\right)\psi_f,\, 2b\psi_f,\, \left(b+\rho\right)\psi_f\right\}.$$

Focusing on the range of reasonable parameter values, $0 < \beta < 1$, $0 < \theta < 1$, $0 \leq \omega < 1$, and $-1 < \rho < 1$, we find that only the RE solution with b^{-} can be E-stable.

B GMM estimation

Moment conditions are obtained using the model's ALM under learning, (25), from which we can define the residual function according to

$$h_t(\Theta) = \pi_t - \beta\rho\phi_{t-1}x_t - \lambda x_t, \tag{51}$$

where Θ denotes the set of parameters requiring estimation. For a given set of instruments, Z_t, the corresponding moment conditions are given by

$$E[Z_t h_t(\Theta)] = 0. \tag{52}$$

The model parameters are then estimated by minimization of the associated GMM objective function

$$g_T\left(\hat{\Theta}\right) = \left[T^{-1}\sum_{t=1}^{T} Z_t h_t\left(\hat{\Theta}\right)\right]' W_T \left[T^{-1}\sum_{t=1}^{T} Z_t h_t\left(\hat{\Theta}\right)\right], \tag{53}$$

which is constructed from the sample counterpart of the moment conditions in (52) and a weighting matrix, W_T. This weighting matrix is optimally defined as a consistent estimator of the inverse of the long-run variance of the moment conditions. Because the variance of (52) depends on the values of Θ, we adopt an iterative GMM estimator (see, e.g., Hall, 2005): we first set $W_T^{(0)} = I$ to obtain the preliminary estimates $\hat{\Theta}^{(0)}$ that minimize (53); we then use the Newey and West (1987) heteroskedasticity and autocorrelation consistent estimator of the variance of the moment conditions evaluated at $\hat{\Theta}^{(0)}$ to obtain a new estimate of $W_T^{(1)}$; we repeat this process until a convergence criterion is achieved.

A numerical optimization routine is used to find the values of $\hat{\Theta}$ that minimize the GMM objective function, (53). For that purpose we adopt a sequential quadratic programming algorithm, namely the 'sqp' option in the fmincon function in Matlab optimization toolbox. The convergence criterion for the iterative estimation of the weighting matrix is based on the Euclidean distance between the successive parameters estimates, i.e., $\left\|\hat{\Theta}^{(i)} - \hat{\Theta}^{(i-1)}\right\| < \epsilon$. In our simulations we set $\epsilon = 10^{-4}$, for which convergence is achieved in about 4 (5) iterations, on average, under $\gamma = 0.02$ ($\gamma = 0.10$).

Whereas the model parameters are reasonably constrained by theory implied boundaries, the parameters associated to the learning algorithm require artificial constrains to avoid numerical instabilities during estimation. Our experimental analysis led us to adopt the following constrains: $0 \leq \gamma \leq 0.5$, $\phi^{REE} - 5 \leq \hat{\phi}_0 \leq \phi^{REE} + 5$, and $0 < \hat{R}_0 \leq 50$. Although these constraints were never violated in the artificial data, the numerical estimation of $\hat{\phi}_0$ and $\hat{R}_0$ often resulted in boundary solutions. These cases are summarized in Table 2, where we observe that increasing the number of estimated parameters tends to increase the number of boundary solutions for the initial estimates. These effects are also amplified when the data true learning gain increases, except for the case of the estimation of the gain itself, where a higher gain facil-

itates estimation. The number of boundary solutions also tends to increase with the sample size under the standard estimation approach, whereas it decreases under the augmented approach. Perhaps the main drawback associated to the augmented approach relates to the estimation of the learning gain, where the estimates turned out unstable for more than 80% of the simulations under the small learning gain and using the small estimations samples.

C Supplementary results

C.1 Validation of analytical MSD expressions

In order to validate the MSD expressions derived in section 3, we ran $100,000$ simulations of the example model of section 2 for different combinations of parameters. For each simulation we draw $10,000 + T$ (pseudo-)random values for u_t from a normal distribution, and use these disturbances to generate artificial observations of y_t and ϕ_t according to the ALM and the learning algorithm specification given by (2) and (6), respectively. We then obtain estimates for $\phi_{t=10,000}$ according to each initialization method described in section 3. The results are presented in table 3 and show a very good fit between our analytical derivations and numerical estimates.

C.2 Initials MSDs with jointly estimated model parameters

Figures 10 and 11 report the initials MSD results for the addition of γ and θ, respectively, to the estimation problem. These results are consistent with our findings discussed on the main text on the basis of Figure 2 for the exercise where both γ and θ are jointly estimated with the learning initial.

C.3 Simulated upper bounds on learning gain for hybrid Phillips curve model

The upper bounds on the learning gain adopted for the empirical exercise are obtained by running 1,000 simulations of the hybrid NKPC model, using the same parametrizations that we are adopting in the estimation, i.e., fixing $\beta = 0.99$ and $\rho = 0.60$, and using RE estimated residuals in order to fix the variances of the residuals needed to simulate the model. The simulations are conducted over a grid of 15 values of $\theta = \{0.01, 0.08, ..., 0.91, 0.99\}$ and 15 values of $\omega = \{0.01, 0.08, ..., 0.91, 0.99\}$, and increasing the learning gain starting from $\gamma = 0.005$ with a step equal to 0.005 until the learning estimates diverge: 1,000 observations are simulated with the first gain, and then if the estimates did not diverge, we move to the next γ, and so on. We check for divergence by looking at the value of the $\hat{b}_t$ learning coefficient estimate, and stop the simulation whenever it goes above the value of 1.5. In general, when the learning estimate

Table 2: Number of simulations where estimates hit lower/upper bounds.

Estimation	Standard Estimation Experiments								Augmented Estimation Experiments							
sample	1	2		3		4			1	2		3		4		
(T)	$\hat{\phi}_0$	$\hat{\phi}_0$	$\hat{\gamma}$	$\hat{\phi}_0$	$\hat{\theta}$	$\hat{\phi}_0$	$\hat{\gamma}$	$\hat{\theta}$	$\hat{\phi}_0$	$\hat{\phi}_0$	$\hat{\gamma}$	$\hat{\phi}_0$	$\hat{\theta}$	$\hat{\phi}_0$	$\hat{\gamma}$	$\hat{\theta}$
Data generated with $\gamma = 0.02$:																
50	0	1030	4151	224	4554	2801	2743	4196	0	462	8137	218	4562	3382	5341	4678
100	0	1060	3028	3	4279	2036	2488	3947	2	156	8326	15	4016	1536	6895	4627
200	0	1080	2007	0	3567	1483	1930	3586	3	46	7487	2	2918	271	7343	4507
500	0	1203	886	0	2467	1465	1004	2883	4	7	3987	5	1216	7	4934	3456
1000	0	1323	284	0	1724	1657	331	2368	0	2	1207	0	340	2	2032	1676
5000	0	1941	0	0	807	2318	0	1352	0	0	1	0	0	0	2	1
Data generated with $\gamma = 0.10$:																
50	149	1436	3287	135	3968	2842	2428	4108	216	903	6211	268	3903	3412	4113	4564
100	157	1599	1548	136	3429	2270	1489	3814	197	609	3881	189	3276	1924	3384	4106
200	197	1787	323	175	2879	2099	503	3384	173	356	1314	160	2331	679	1569	3080
500	335	2163	5	297	2062	2339	47	2858	102	142	135	103	997	144	200	1275
1000	479	2540	0	454	1372	2695	2	2562	57	71	14	62	288	70	21	354
5000	1282	3644	0	1092	922	3443	0	2387	16	15	0	17	0	16	0	1

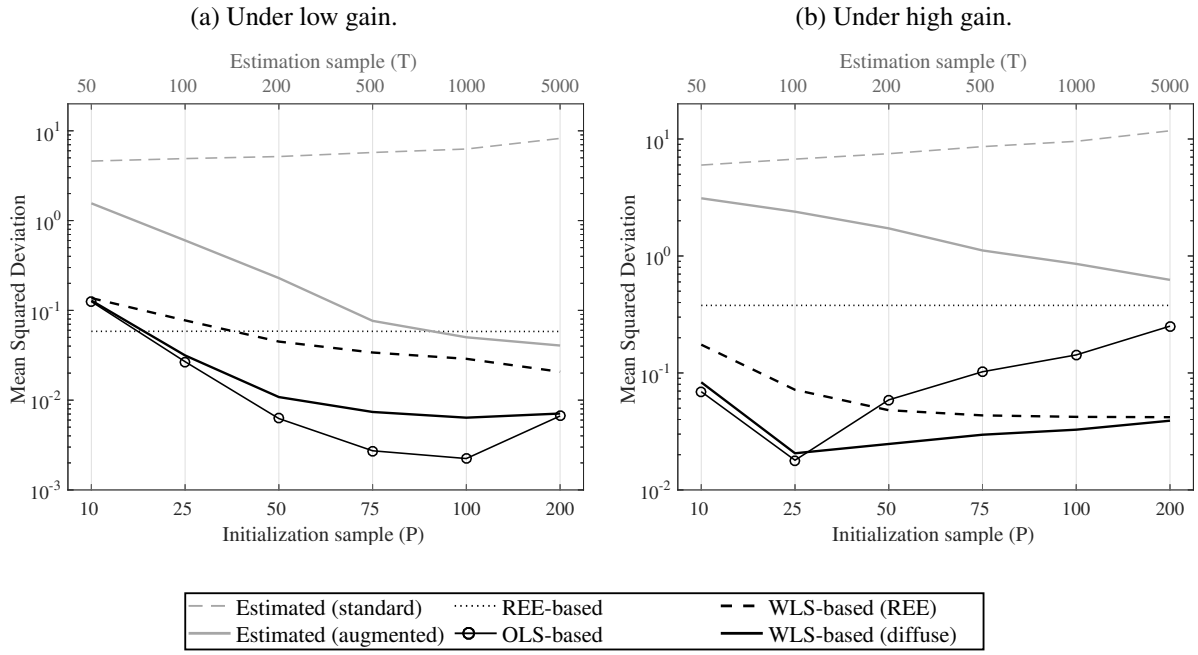
Notes: Statistics based on 10,000 simulations of the benchmark Phillips curve model under learning.

Table 3: Analytical and numerical MSDs for initials obtained under the example model.

Method	$\sigma_u^2 = 1$				$\sigma_u^2 = 4$			
	$\beta = 0.9$		$\beta = -5$		$\beta = 0.9$		$\beta = -5$	
$\gamma \Rightarrow$	0.02	0.25	0.02	0.25	0.02	0.25	0.02	0.25
REE-based:								
	0.100	1.269	0.002	0.083	0.400	5.082	0.007	0.331
	(0.100)	(1.266)	(0.002)	(0.083)	(0.400)	(5.063)	(0.007)	(0.333)
WLS-based:								
$P = 10$	0.067	0.004	0.001	0.000	0.267	0.016	0.005	0.001
	(0.067)	(0.004)	(0.001)	(0.000)	(0.267)	(0.016)	(0.005)	(0.001)
$P = 25$	0.037	0.000	0.001	0.000	0.145	0.000	0.003	0.000
	(0.036)	(0.000)	(0.001)	(0.000)	(0.146)	(0.000)	(0.003)	(0.000)
$P = 50$	0.013	0.000	0.000	0.000	0.053	0.000	0.001	0.000
	(0.013)	(0.000)	(0.000)	(0.000)	(0.053)	(0.000)	(0.001)	(0.000)
$P = 100$	0.002	0.000	0.000	0.000	0.007	0.000	0.000	0.000
	(0.002)	(0.000)	(0.000)	(0.000)	(0.007)	(0.000)	(0.000)	(0.000)
OLS-based:								
$P = 10$	0.079	0.066	0.051	0.045	0.317	0.261	0.205	0.180
	(0.079)	(0.065)	(0.051)	(0.045)	(0.316)	(0.261)	(0.204)	(0.182)
$P = 25$	0.022	0.284	0.008	0.063	0.089	1.128	0.033	0.252
	(0.022)	(0.283)	(0.008)	(0.063)	(0.089)	(1.132)	(0.032)	(0.253)
$P = 50$	0.006	0.580	0.002	0.072	0.025	2.304	0.007	0.288
	(0.006)	(0.577)	(0.002)	(0.072)	(0.025)	(2.308)	(0.007)	(0.290)
$P = 100$	0.003	0.905	0.001	0.077	0.013	3.609	0.003	0.308
	(0.003)	(0.903)	(0.001)	(0.078)	(0.013)	(3.612)	(0.003)	(0.311)
SSR estimation-based:								
$T = 10$	0.147	0.543	0.005	0.017	0.591	2.152	0.019	0.070
	(0.147)	(0.542)	(0.005)	(0.018)	(0.588)	(2.167)	(0.019)	(0.070)
$T = 100$	0.050	0.541	0.002	0.017	0.199	2.146	0.006	0.070
	(0.050)	(0.540)	(0.002)	(0.018)	(0.199)	(2.160)	(0.006)	(0.070)
$T = 250$	0.049	0.541	0.002	0.017	0.196	2.146	0.006	0.070
	(0.049)	(0.540)	(0.002)	(0.018)	(0.196)	(2.160)	(0.006)	(0.070)
$T = 1000$	0.049	0.541	0.002	0.017	0.196	2.146	0.006	0.070
	(0.049)	(0.540)	(0.002)	(0.018)	(0.196)	(2.160)	(0.006)	(0.070)

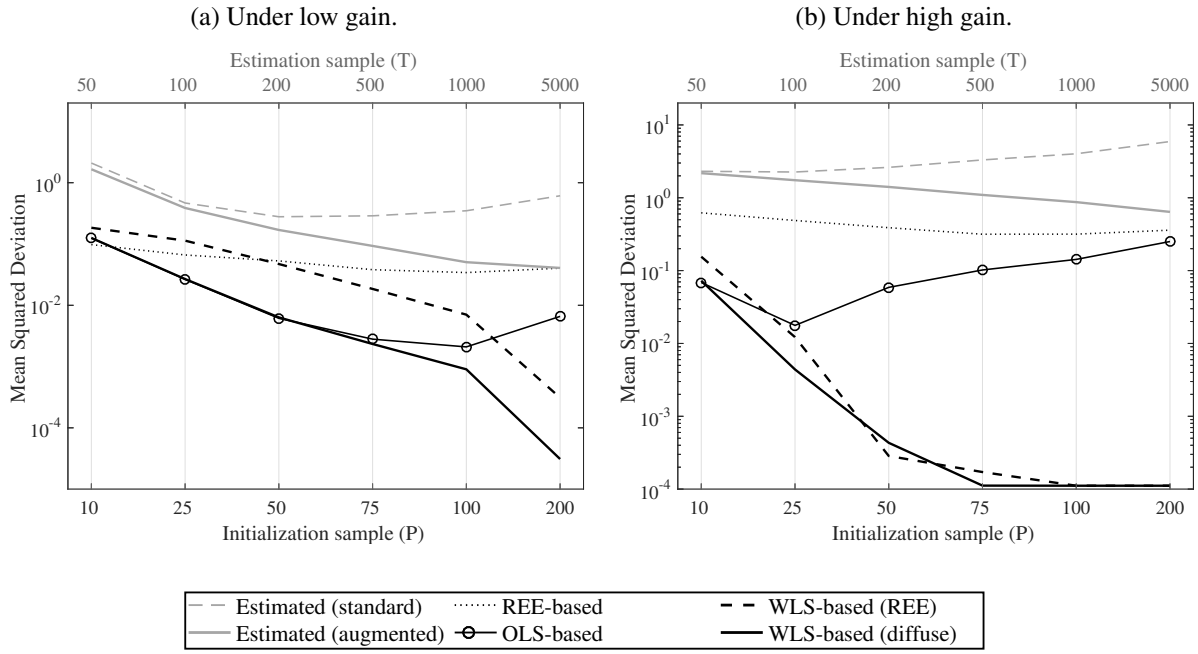
Notes: All statistics refer to MSD estimates. Values in parentheses come from the analytical expressions derived in the text, whereas those without parentheses come from the numerical simulation.

Figure 10: Mean Squared Deviations of initials with ϕ_0 and γ jointly estimated.



Notes: See notes of Figure 1. The estimates of γ used for the WLS-based initials are those obtained under the augmented estimation approach with $T = 200$.

Figure 11: MSDs - ϕ_0 and θ jointly estimated.



Notes: See notes of Figure 1. The estimates of θ used for the REE-based initials are those obtained under the augmented estimation approach.

reached that value, the simulated inflation series was already exploding. The upper bounds are then obtained by averaging the maximum gain values across the 1,000 simulations for each combination of parameters. For the empirical estimation of the model, we obtain an interpolant function, $G(\theta, \omega)$, using cubic splines based on the gridded gain upper bounds presented in Figure 7. This function is then supplied to the numerical estimation routine as a nonlinear restriction on the estimates.

C.4 Empirical estimates for alternative sample

Estimates of macroeconomic models may be sensitive to the historical period underlying the sample used for estimation. U.S. post-WWII inflation history over our sample of data has been marked by two main episodes of interest: (i) the Great Inflation period, starting around early-1970s and ending by mid-1980s; and, (ii) the more recent period that followed the 2007-08 financial crisis. To check for the sensitiveness of our results, we re-estimated the model discussed in the text by focusing on samples that exclude these periods. Results are presented in Tables 4 and 5, the former excluding the Great Inflation period and the latter also excluding the period since the financial crisis. Finally, we also re-estimated the model with an earlier sample beginning in order to check for the sensitiveness of the implied inflation expectations at the start of the Great Inflation period. These results are presented in Table 6.

Table 4: Empirical estimates of US NKPC - 1985q1 - 2016q3.

Exercise	Parameter estimates: Structural		Learning				Implied reduced form parameters		
	θ	ω	γ	a_0	b_0	c_0	ψ_f	ψ_b	δ
Rational expectations									
	0.86	**0.37**					0.70	0.30	0.01
	(0.11)	**(0.06)**							
Learning with jointly estimated initials									
- Standard	0.87	0.92	**0.05**	**3.29**	-0.02	**-3.73**	0.48	0.52	0.00
	(3.21)	(3.39)	**(0.02)**	**(1.16)**	(0.52)	**(1.13)**			
- Augmented	0.99	0.09	0.00	1.12	-1.10	-0.35	0.91	0.08	0.00
	(7.07)	(1.18)	(0.01)	(0.79)	(0.68)	(0.70)			
Learning with REE-based initials									
- Standard	**0.77**	**0.17**	**0.08**	0.00	0.22	0.15	0.81	0.18	0.05
	(0.10)	**(0.08)**	**(0.02)**						
- Augmented	**0.65**	**0.30**	**0.06**	0.00	0.46	0.33	0.68	0.32	0.09
	(0.08)	**(0.04)**	**(0.02)**						
Under learning with WLS-based initials									
- Standard	**0.47**	**0.37**	0.07	0.30	0.77	-0.04	0.56	0.44	0.21
	(0.03)	**(0.05)**	(0.04)						
- Augmented	**0.66**	**0.53**	0.00	0.31	0.81	0.02	0.55	0.45	0.05
	(0.11)	**(0.06)**	(0.00)						
Under learning with OLS-based initials									
- Standard	0.99	0.01	**0.09**	0.31	0.81	0.02	0.98	0.01	0.00
	(2.48)	(0.17)	**(0.03)**						
- Augmented	**0.64**	**0.53**	0.01	0.31	0.81	0.02	0.54	0.45	0.05
	(0.08)	**(0.07)**	(0.01)						

Notes: See notes to Table 1.

Table 5: Empirical estimates of US NKPC - 1985q1 - 2007q4.

Exercise	Parameter estimates: Structural		Learning				Implied reduced form parameters		
	θ	ω	γ	a_0	b_0	c_0	ψ_f	ψ_b	δ
Rational expectations									
	0.91	**0.20**					0.81	0.18	0.01
	(0.13)	**(0.06)**							
Learning with jointly estimated initials									
- Standard	0.59	0.10	**0.01**	**1.19**	**-0.74**	**6.70**	0.85	0.15	0.22
	(0.38)	(0.14)	**(0.00)**	**(0.14)**	**(0.09)**	**(1.01)**			
- Augmented	**0.33**	**0.57**	**0.04**	**4.43**	-0.39	**-3.77**	0.36	0.63	0.22
	(0.05)	**(0.07)**	**(0.01)**	**(0.99)**	(0.25)	**(1.43)**			
Learning with REE-based initials									
- Standard	0.95	0.17	**0.08**	0.00	0.18	0.01	0.84	0.15	0.00
	(0.56)	(0.11)	**(0.02)**						
- Augmented	0.69	0.25	**0.07**	0.00	0.35	0.26	0.73	0.27	0.08
	(0.09)	(0.06)	**(0.02)**						
Under learning with WLS-based initials									
- Standard	**0.66**	**0.51**	0.00	0.31	0.81	0.02	0.56	0.44	0.05
	(0.22)	**(0.14)**	(0.01)						
- Augmented	**0.66**	**0.31**	0.06	0.31	0.78	-0.04	0.68	0.32	0.08
	(0.06)	**(0.07)**	(0.04)						
Under learning with OLS-based initials									
- Standard	0.97	0.94	0.00	0.31	0.78	-0.04	0.51	0.49	0.00
	(15.87)	(15.78)	(0.01)						
- Augmented	**0.66**	**0.34**	0.06	0.31	0.78	-0.04	0.65	0.34	0.08
	(0.06)	**(0.06)**	(0.03)						

Notes: See notes to Table 1.

Table 6: Empirical estimates of US NKPC - 1966q4 - 2016q3.

Exercise	Parameter estimates: Structural		Learning				Implied reduced form parameters		
	θ	ω	γ	a_0	b_0	c_0	ψ_f	ψ_b	δ
Rational expectations									
	0.84	**0.40**					0.67	0.32	0.01
	(0.07)	**(0.06)**							
Learning with jointly estimated initials									
- Standard	0.96	0.29	0.06	1.45	0.11	0.56	0.76	0.23	0.00
	(1.66)	(0.70)	(0.03)	(2.01)	(1.71)	(0.81)			
- Augmented	**0.72**	0.22	**0.07**	4.89	-1.16	0.03	0.76	0.23	0.07
	(0.06)	(0.17)	**(0.02)**	(3.03)	(0.60)	(2.21)			
Learning with REE-based initials									
- Standard	**0.78**	0.04	**0.10**	0.00	0.05	0.15	0.94	0.05	0.06
	(0.07)	(0.09)	**(0.02)**						
- Augmented	**0.72**	**0.38**	**0.05**	0.00	0.53	0.17	0.65	0.35	0.05
	(0.11)	**(0.07)**	**(0.01)**						
Under learning with WLS-based initials									
- Standard	**0.77**	0.03	**0.10**	0.22	0.56	-0.06	0.95	0.04	0.07
	(0.06)	(0.09)	**(0.02)**						
- Augmented	**0.85**	**0.35**	**0.05**	0.23	0.50	-0.04	0.70	0.29	0.01
	(0.17)	**(0.06)**	**(0.02)**						
Under learning with OLS-based initials									
- Standard	**0.77**	0.01	**0.11**	0.23	0.50	-0.04	0.98	0.01	0.07
	(0.06)	(0.09)	**(0.02)**						
- Augmented	**0.88**	**0.32**	**0.06**	0.23	0.50	-0.04	0.73	0.27	0.01
	(0.21)	**(0.08)**	**(0.01)**						

Notes: See notes to Table 1.

References

Barucci, E. and L. Landi (1997). Least mean squares learning in self-referential linear stochastic models. *Economics Letters 57*(3), 313–317.

Berardi, M. and J. K. Galimberti (2013). A note on exact correspondences between adaptive learning algorithms and the kalman filter. *Economics Letters 118*(1), 139–142.

Berardi, M. and J. K. Galimberti (2014). A note on the representative adaptive learning algorithm. *Economics Letters 124*(1), 104 – 107.

Berardi, M. and J. K. Galimberti (2016, December). On the Initialization of Adaptive Learning in Macroeconomic Models. Technical report, KOF Working Papers, Zürich.

Berardi, M. and J. K. Galimberti (2017). Smoothing-based initialization for learning-to-forecast algorithms. *Macroeconomic Dynamics in press*.

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

Bullard, J. and S. Eusepi (2005). Did the great inflation occur despite policymaker commitment to a taylor rule? *Review of Economic Dynamics 8*(2), 324 – 359.

Carboni, G. and M. Ellison (2009). The great inflation and the greenbook. *Journal of Monetary Economics 56*(6), 831 – 841.

Carceles-Poveda, E. and C. Giannitsarou (2007). Adaptive learning in practice. *Journal of Economic Dynamics and Control 31*(8), 2659–2697.

Carceles-Poveda, E. and C. Giannitsarou (2008). Asset pricing with adaptive learning. *Review of Economic Dynamics 11*(3), 629 – 651.

Chevillon, G., M. Massmann, and S. Mavroeidis (2010). Inference in models with adaptive learning. *Journal of Monetary Economics 57*(3), 341–351.

Eusepi, S. and B. Preston (2011, October). Expectations, learning, and business cycle fluctuations. *American Economic Review 101*(6), 2844–72.

Evans, G. W. and S. Honkapohja (1998). Stochastic gradient learning in the cobweb model. *Economics Letters 61*(3), 333–337.

Evans, G. W. and S. Honkapohja (2001). *Learning and expectations in macroeconomics*. Frontiers of Economic Research. Princeton, NJ: Princeton University Press.

Evans, G. W. and S. Honkapohja (2009, 06). Robust Learning Stability with Operational Monetary Policy Rules. In K. Schmidt-Hebbel, C. E. Walsh, N. L. S. Editor), and K. S.-H. (Series (Eds.), *Monetary Policy under Uncertainty and Learning*, Volume 13 of *Central Banking, Analysis, and Economic Policies Book Series*, Chapter 5, pp. 145–170. Central Bank of Chile.

Fuhrer, J. C., G. R. Moore, and S. D. Schuh (1995). Estimating the linear-quadratic inventory model maximum likelihood versus generalized method of moments. *Journal of Monetary Economics 35*(1), 115 – 157.

Gali, J. and M. Gertler (1999, October). Inflation dynamics: A structural econometric analysis. *Journal of Monetary Economics, Elsevier 44*(2), 195–222.

Gaus, E. and S. Ramamurthy (2014, August). Estimation of constant gain learning models. Working Papers 12-01, Ursinus College, Department of Economics.

Hall, A. R. (2005). *Generalized Method of Moments*. Advanced Texts in Econometrics. Oxford University Press.

Huang, K. X., Z. Liu, and T. Zha (2009). Learning, adaptive expectations and technology shocks. *The Economic Journal 119*(536), 377–405.

Ljung, L. and T. Soderstrom (1983). *Theory and Practice of Recursive Identification*. The MIT Press.

Lubik, T. A. and C. Matthes (2014, January). Indeterminacy and learning: An analysis of monetary policy in the great inflation. Working Paper 14-2, Federal Reserve Bank of Richmond.

Marcet, A. and J. P. Nicolini (2003). Recurrent hyperinflations and learning. *American Economic Review 93*(5), 1476–1498.

Mavroeidis, S., M. Plagborg-Møller, and J. H. Stock (2014). Empirical evidence on inflation expectations in the new keynesian phillips curve. *Journal of Economic Literature 52*(1), 124–88.

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

Milani, F. (2008). Learning, monetary policy rules, and macroeconomic stability. *Journal of Economic Dynamics and Control 32*(10), 3148 – 3165.

Milani, F. (2011). Expectation shocks and learning as drivers of the business cycle. *The Economic Journal 121*(552), 379–401.

Milani, F. (2014). Learning and time-varying macroeconomic volatility. *Journal of Economic Dynamics and Control 47*(0), 94 – 114.

Newey, W. K. and K. D. West (1987, May). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–08.

Ormeño, A. and K. Molnár (2015, June). Using survey data of inflation expectations in the estimation of learning and rational expectations models. *Journal of Money, Credit and Banking 47*(4), 673–699.

Orphanides, A. and J. C. Williams (2005a, November). 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.

Pfajfar, D. and E. Santoro (2010). Heterogeneity, learning and information stickiness in inflation expectations. *Journal of Economic Behavior & Organization 75*(3), 426–444.

Primiceri, G. E. (2006). Why inflation rose and fell: Policy-makers' beliefs and u. s. postwar stabilization policy. *The Quarterly Journal of Economics 121*(3), 867–901.

Sargent, T., N. Williams, and T. Zha (2006). Shocks and government beliefs: The rise and fall of american inflation. *American Economic Review 96*(4), 1193–1224.

Sargent, T. J. (1999). *The Conquest of American Inflation*. Princeton, NJ: Princeton University Press.

Slobodyan, S. and R. Wouters (2012a). Learning in a medium-scale dsge model with expectations based on small forecasting models. *American Economic Journal: Macroeconomics 4*(2), 65–101.

Slobodyan, S. and R. Wouters (2012b). Learning in an estimated medium-scale dsge model. *Journal of Economic Dynamics and Control 36*(1), 26 – 46.

Williams, N. (2003, January). Adaptive learning and business cycles. Mimeo.