Economics Working Paper Series
Faculty of Business, Economics and Law, AUT

Initial Beliefs Uncertainty and Information Weighting in the Estimation of Models with Adaptive Learning

Jaqueson K. Galimberti

2021/01

Initial Beliefs Uncertainty and Information Weighting in the Estimation of Models with Adaptive Learning

Jaqueson K. Galimberti*

Abstract

This paper evaluates how the way agents weight information when forming expectations can affect the econometric estimation of models with adaptive learning. One key new finding is that misspecification of the uncertainty about initial beliefs under constant-gain least squares learning can generate a time-varying profile of weights given to past observations, distorting the estimation and behavioural interpretation of this mechanism in small samples of data. This result is derived under a new representation of the learning algorithm that penalizes the effects of misspecification of the learning initials. Simulations of a forward-looking Phillips curve model with learning indicate that (i) misspecification of initials uncertainty can lead to substantial biases to estimates of expectations relevance for inflation, and (ii) that these biases can spill over to estimates of inflation rates responsiveness to output gaps. An empirical application with U.S. data shows the relevance of these effects.

Keywords: expectations, adaptive learning, bounded rationality, macroeconomics.

JEL codes: E70, D83, D84, D90, E37, C32, C63.

"The longer you can look back,
the farther you can look forward."
–Winston Churchill

1 Introduction

Adaptive learning can generate out-of-equilibrium expectations that help explain deviations from rational expectations and an economy's transitional dynamics towards equilibrium. Under adaptive learning, agents' beliefs are modelled through the assumption of a recursive learning mechanism that updates agents' perceptions about the economy as new data observations become available. The weights given to these observations are key determinants of the degree

*Auckland University of Technology, New Zealand; Centre for Applied Macroeconomic Analysis, Australian National University; and KOF Swiss Economic Institute, ETH Zurich. Corresponding author: jaqueson.galimberti@aut.ac.nz, https://sites.google.com/site/jkgeconoeng/.

of persistence introduced by adaptive learning in the evolution of expectations, and, hence, are important factors in the explanation of deviations from rational expectations predictions. Besides, initial beliefs and an estimate of agents' uncertainty about those beliefs need to be specified and can account for some of the explanatory power of learning. In this paper I study the implications of alternative specifications of the learning mechanism regarding how new information is weighted relative to assumed initial beliefs, and the effects of these assumptions on the estimation of models with adaptive learning.

I focus on applications with a recursive least squares (RLS), a popular learning mechanism assumed to represent agents' econometric learning in the bounded rationality literature. One contribution of this paper is the description of a renewed and more general non-recursive representation of the RLS. Namely, I show that the RLS is more properly represented by a penalized weighted least squares (WLS) estimator, where a penalty term accounts for the effects of the learning initial estimates. The pace of RLS learning is regulated by a sequence of learning gains, which also determines how different pieces of information are weighted in the implied estimates of agents' perceived law of motion. The framework proposed in this paper provides flexible analytical expressions for the calculation of these weights under alternative assumptions on the behaviour of the learning gains, including the constant-gain (CG) specification that, since Sargent (1999), has received great attention in applied research due to its capability to generate perpetual learning.[1] Importantly, it is well known that the influence of the learning initials can become non-negligible in applications with a CG, where the weights given to past observations decrease geometrically (see, e.g., Carceles-Poveda and Giannitsarou, 2007).

In fact, the main contribution of this paper is the use of this framework to derive an important result regarding the distortionary effects of misspecified initials uncertainty under CG learning. From a model estimation point of view, "initials uncertainty" is a parameter determining the confidence agents are assumed to have on their beliefs at the beginning of the econometrician's estimation sample. Intriguingly, I find that the weighting ascribed to new data information into agents' beliefs can be substantially affected by the uncertainty assumed around agents' initial beliefs. Particularly, I find that the assumption of a *diffuse initial*, representing a situation where the agent is completely uncertain about his/her initial beliefs, implies that the profile of weights given to past observations under CG learning is in fact time-varying, causing a geometric decay of weighting stronger than would have been expected from asymptotic analysis of the learning algorithm. In other words, the use of diffuse initials is equivalent to the use of higher starting learning gains that decrease as the sample grows and only converges to the actual CG asymptotically.

I show that this result can have important implications for the estimation of models with learning, as it can lead to an overweighting of the initial sample of observations used for model

[1] Other specifications that can be easily reproduced in this paper's framework include the traditional decreasing-gain (Marcet and Sargent, 1989), as well as more sophisticated mechanisms such as endogenous gain-switching (Marcet and Nicolini, 2003) and age-dependent (Malmendier and Nagel, 2016) gain specifications.

estimation. Without a proper account for the learning initial, the estimation of models under the assumption of a CG over increasing samples of data would imply agents give a decreasing weight to more recent observations. In other terms, as we accumulate more data about the economy, and use this additional data in the estimation of models, the underlying estimated expectations are likely to become less sensitive to new information than they used to be with the earlier, and hence smaller, samples of data. This can introduce a downward bias on renewed estimates of the relevance of the expectations formation mechanism in the determination of the latest economic developments.

On the other hand, diffuse initials can be used as a way to speed up convergence of learning estimates. In most applications, learning is assumed to represent a process that was ongoing prior to the beginning of the econometrician's estimation sample. In the lack of proper estimates of such initial beliefs, diffuse initials offer an interesting alternative to be used in training samples. Hence, the effect of misspecified initials uncertainty I discuss in this paper is a concern mainly for its use within the model estimation sample. Considering that this approach has been considered in previous applications in the literature (e.g., Slobodyan and Wouters, 2012; Markiewicz and Pick, 2014; Lubik and Matthes, 2016), it is important to understand the effects that the associated information weighting distortions can have on model estimates.

To quantify these potential biases I simulate the estimation of a new Keynesian Phillips curve model and find that, indeed, diffuse initials lead to stronger small sample distortions in the model estimates. Particularly, the misspecified initials result in a systematic underestimation of the relevance of expectations in this model, as well as an overestimation of the responsiveness of inflation rates to measures of economic slack. The simulation analysis also allows an investigation of the channels through which these effects emerge, pointing to an increase in the variance of expectations associated with the diffuse initials as the main cause for the estimation biases. This result is consistent with the analysis of information weighting of CG learning using a diffuse initial; particularly, the equivalent higher gains at the beginning of the estimation sample lead to higher variability in the learning estimates, which ultimately increases the variance of the implied expectations. Empirical estimates with decadal sub-samples of US data also indicate that the diffuse initials can distort estimates of the relevance of expectations for the determination of inflation, particularly implying a lower degree of violations to expectational stability conditions across the different sub-samples.

The remainder of this paper is split into four other sections: Section 2 outlines the learning framework and derives the main theoretical results about information weighting and initials uncertainty; Section 3 presents simulation analysis of the distortionary effects of diffuse initials in the estimation of a forward-looking Phillips curve model with CG learning; Section 4 presents an empirical application of the same model to U.S. data; Section 5 concludes with some remarks. Proofs and detailed derivations are provided in the Appendix.

2 Learning Framework

In this section I outline the general framework of recursive learning in order to derive the information weighting implications of different specifications of the learning gains and initials. Focusing on the case of a constant-gain, I then show how the specification of uncertainty about the initial learning estimates can distort the profile of weights that this popular learning mechanism assigns to observations at the beginning of estimation samples.

2.1 Background

In models with adaptive learning a perceived law of motion (PLM) is specified relating the variables agents are assumed to observe and those variables they care and need to form expectations about. Focusing on a univariate case, a typical PLM specification is given by a linear regression model of the form

$$y_t = \mathbf{x}_t' \phi_t + \varepsilon_t, \tag{1}$$

where y_t is assumed to be related to a vector of (pre-determined) variables, $\mathbf{x}_t = (x_{1,t}, \ldots, x_{k,t})'$, through the vector of (possibly time-varying) coefficients $\phi_t = (\phi_{1,t}, \ldots, \phi_{k,t})'$, and ε_t denotes a white noise disturbance term. This specification can be easily extended to a multivariate context, by augmenting y_t, ε_t, and ϕ_t with extra columns, and to different specifications of lag/lead in the timing of expectations, by adjusting the timing of $\mathbf{x}_t$ elements.

In a typical economic modelling context, the observations of y_t needed to estimate (1), as well as some or all of the regressors in $\mathbf{x}_t$, are endogenously determined within a hypothetical structural model. These observations are the result of market equilibrium and the interaction between the economic decisions by the many different actors that compose the economy, such as households, firms, and policymakers. Hence, due to the relevance of expectations for these agents' economic decisions, the same macroeconomic outcomes that are relevant for the formation of expectations are themselves determined by expectations, a feature often called self-referentiality.

Notwithstanding, for the purposes of deriving weighting expressions I will abstract from one side of this self-referential nature of expectations, and focus instead on the modelling of agents' PLM. I.e., I will assume agents form expectations according to equation (1) without accounting for the endogeneity of the involved variables. Notice, as is usually the case in applications of adaptive learning, this implies some degree of bounded rationality in the way agents form expectations. The effects of self-referentiality are taken into account in the simulation and empirical estimation exercises presented in later sections.

2.2 Learning and initials uncertainty

A recursive estimator is assumed to represent how agents update their PLM estimates as new observations become available. One popular algorithm is given by the Recursive Least Squares (RLS),

$$\phi_t = \phi_{t-1} + \gamma_t \mathbf{R}_t^{-1} \mathbf{x}_t \left(y_t - \mathbf{x}_t' \phi_{t-1} \right), \tag{2}$$

$$\mathbf{R}_t = \mathbf{R}_{t-1} + \gamma_t \left(\mathbf{x}_t \mathbf{x}_t' - \mathbf{R}_{t-1} \right), \tag{3}$$

where $\mathbf{R}_t$ stands for an estimate of regressors' matrix of second moments, $E\left[\mathbf{x}_t\mathbf{x}_t'\right]$, and γ_t is a learning gain parameter. The learning gain is an important parameter of the learning mechanism because it determines how quickly new information is incorporated into the recursive estimates, and hence, how quickly agents react to different pieces of information (this relation will be discussed in the next subsection). Moreover, as a recursive process, these estimates need to initialized: ϕ_0 are initial estimates representing agents' beliefs at the beginning of the econometrician's sample of data, and the inverse of $\mathbf{R}_0$ can be interpreted as a measure of the uncertainty agents assign to these initial estimates.

The main focus of this paper is about the determination of the initial beliefs uncertainty, $\mathbf{R}_0$, in a context of econometric estimation of economic models with learning. Naturally, the initial estimates should ideally be set or estimated to be consistent with plausible agents' beliefs at the beginning of the modelled sample. Berardi and Galimberti (2017b) study methods for the estimation of ϕ_0 aimed to achieve such a goal, although assuming a fixed $\mathbf{R}_0$ across model estimation exercises. As it turns out, alternative assumptions of $\mathbf{R}_0$ can play an important role in the estimation of models with learning. The main contribution of this paper is to provide an analysis of this component.

From a Bayesian point of view, $\mathbf{R}_t$ is inversely related to the uncertainty in the corresponding Kalman filter estimates of ϕ_t modelled as a random walk (see Evans et al., 2010; Berardi and Galimberti, 2013). Hence, $\mathbf{R}_0 \to \mathbf{0}$, henceforth denoted as diffuse initials, can be interpreted as increasing the uncertainty about the initial estimates, in which case the observations at the beginning of the estimation sample will be given extra weight to compensate for the initials uncertainty.[2]

This effect has two main implications for the estimation of models with learning. First, diffuse initials can be used as a way to accelerate convergence of learning estimates to a process representing ongoing learning that was already happening prior to the beginning of the econometrician's estimation sample. This particular property makes the diffuse initials an interesting alternative to be used in training samples. Second, within an estimation sample, the overweighting of initial observations distorts the representativeness of the implied expectations, which, in

[2] Also, notice that if $\mathbf{R}_0 = \mathbf{0}$ (exactly rather than as a limit), (2)-(3) implies that $\hat{\phi}_1 = \left(\mathbf{x}_1\mathbf{x}_1'\right)^{-1}\mathbf{x}_1 y_1$, which will be indeterminate for $k > 1$. For this reason, in the applications that follow I approximate diffuse initials by downscaling a reference $\mathbf{R}_0$ towards zero by multiplying it by a small constant.

turn, can affect model estimates that depend on the behaviour of such expectations. To be more precise, in what follows I show how information weighting can be traced back to the joint definition of the learning gains and initials using a renewed and more general non-recursive representation of the learning algorithm.

2.3 Non-recursive form and information weighting

The weight given to a sample observation determines the amount of information from that particular observation that is incorporated into the PLM estimates. In the RLS algorithm of equations (2)-(3), such weighting of information is controlled by the sequence of learning gains. More precisely, the sequence of learning gains can be related with the relative weights given to sample observations in the estimation process. In order to draw this relationship it is useful to consider the non-recursive formulation corresponding to this estimation problem. When initialized from arbitrary initials, $\tilde{\phi}_0$ and $\tilde{\mathbf{R}}_0$, the RLS has a non-recursive form given by[3]

$$\begin{aligned}\phi_t &= \arg\min_{\tilde{\phi}_t} \sum_{i=1}^{t} \omega_{t,i} \left(y_i - \mathbf{x}_i'\tilde{\phi}_t\right)^2 + \omega_{t,0}\left(\tilde{\phi}_0' - \tilde{\phi}_t'\right)\tilde{\mathbf{R}}_0\left(\tilde{\phi}_0 - \tilde{\phi}_t\right), && (4)\\ &= \left[\sum_{i=1}^{t} \omega_{t,i}\mathbf{x}_i\mathbf{x}_i' + \omega_{t,0}\tilde{\mathbf{R}}_0\right]^{-1}\left[\sum_{i=1}^{t} \omega_{t,i}\mathbf{x}_i y_i + \omega_{t,0}\tilde{\mathbf{R}}_0\tilde{\phi}_0\right], && (5)\end{aligned}$$

where the weights are related to the sequence of learning gains according to

$$\omega_{t,i} = \begin{cases} \prod_{j=1}^{t}\left(1-\gamma_j\right) & for\ i = 0\,(initial), \\ \gamma_i \prod_{j=i+1}^{t}\left(1-\gamma_j\right) & for\ 0 < i < t, \\ \gamma_t & for\ i = t. \end{cases} \quad (6)$$

Thus, when the initials are taken into account, the RLS is equivalent to a Weighted Least Squares (WLS) estimation problem augmented with a penalty on squared deviations between estimates and initials. To the best of my knowledge this non-recursive formulation of the RLS for arbitrary initials has never been outlined in the previous literature. In fact, the origins of the RLS can be traced back as the recursive formulation of the WLS solution (without the penalty on initials) to the minimization of the sum of weighted error squares in the systems identification literature (see, e.g., Ljung and Soderstrom, 1983). Hence, the innovation here stems from following the inverse approach, i.e., taking the recursive form of (2)-(3) with initials $\{\tilde{\phi}_0, \tilde{\mathbf{R}}_0\}$ as the starting point, I obtain (5)-(6), which, in turn, can be translated as a solution to the estimation problem in equation (4).

The non-recursive formulation above allows the calculation of such weights for any arbitrary sequence of learning gains. Also notice that the weights, $\omega_{t,i}$, defined in equation (6), are

[3] See Appendix A.1 for the proof.

already in relative terms, as obtained by dividing the absolute weights by the sum of weights given to sample observations and the initial. This follows from the fact that, under the correspondence between the RLS and the penalized WLS outlined in this paper, the sum of weights will always be equal to unity.[4]

Moreover, it is often interesting to evaluate how the observation weights evolve relative to the last observation in the sample, i.e., redefining equation (6) in terms of lags relative to the end of the sample, $\varpi_{t,l} = \omega_{t,t-l}$, in which case we look at

$$\varpi_{t,l} = \begin{cases} \prod_{j=1}^{t}\left(1-\gamma_j\right) & for\; l = t, \\ \gamma_{t-l}\prod_{j=0}^{l-1}\left(1-\gamma_{t-j}\right) & for\; 0 < l < t, \\ \gamma_t & for\; l = 0. \end{cases} \tag{7}$$

2.4 Constant-gain learning and diffuse initials

The constant-gain (CG) learning specification was introduced in the applied learning literature by Evans and Honkapohja (1993) and became popular after Sargent (1999) for its improved capability of tracking the evolution of time-varying environments. This specification has also been under the spotlight of the most recent research on the dynamic modelling of expectations for its potential of generating escape dynamics over finite stretches of time (see Williams, 2019) and asymptotically stable distributions of beliefs (see Galimberti, 2019).

One important property of the CG-RLS relates to the persistent influence of the learning initials. Under CG-RLS learning, $\gamma_t^{cg} = \bar{\gamma}$, and the weights are given by

$$\omega_{t,i}^{cg} = \begin{cases} (1-\bar{\gamma})^t & for\; i = 0, \\ \bar{\gamma}(1-\bar{\gamma})^{t-i} & for\; 0 < i \leq t. \end{cases}$$

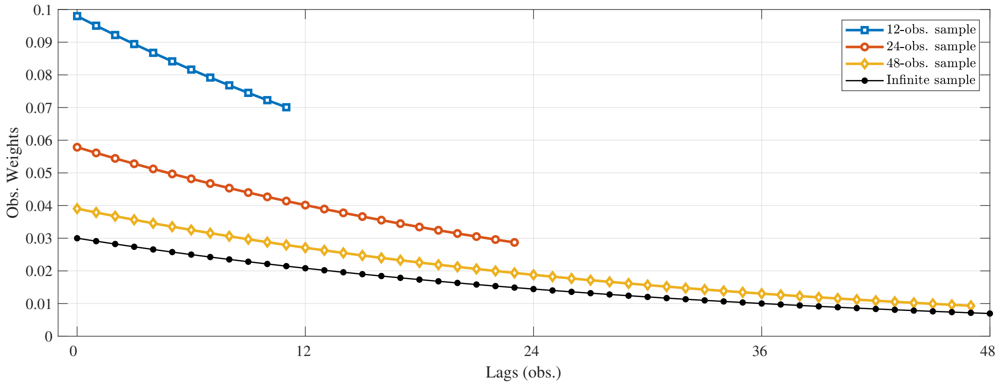
Hence, the relative weights given to sample information by the CG-RLS decrease with the observation lag ($l = t - i$), while the weight given to the initial decreases with the sample size. However, the duration of the effects of the initials within finite samples will depend on the learning gain. For example, the number of observations needed to equate the cumulative weight given to sample information to the weight given to the initials can be easily calculated as

$$\sum_{j=1}^{i^*} \omega_{t,j}^{cg} = \omega_{t,0}^{cg},$$

$$i^* = \frac{\log\left(1/2\right)}{\log\left(1-\bar{\gamma}\right)}.$$

[4] See Appendix A.2 for the proof.

Figure 1: Constant-gain weights under diffuse initials.



Notes: Weights calculated using equation (8) with $\bar{\gamma} = 0.03$.

For a learning gain of $\bar{\gamma} = 0.03$, a value typically found in applications with quarterly macroeconomic data (see Berardi and Galimberti, 2017a, for a review), $i^* \simeq 23$, i.e., it takes about six years of quarterly data for the CG-RLS to assign a higher weight to the sample of observations than the weight given to the learning initials in the PLM estimates. This clearly highlights the importance of properly estimating such learning initials under CG learning.

Another important property of the CG-RLS relates to its asymptotic weighting behaviour relative to lagged observations. Although the CG-RLS assigns a vanishing weight to any given sample observation i, the weight given to observations at a fixed lag l do not change with t. This property makes the CG-RLS particularly well suited for modelling the behavioural assumption that agents give a higher emphasis to the more recent observations than to those collected farther into the past.

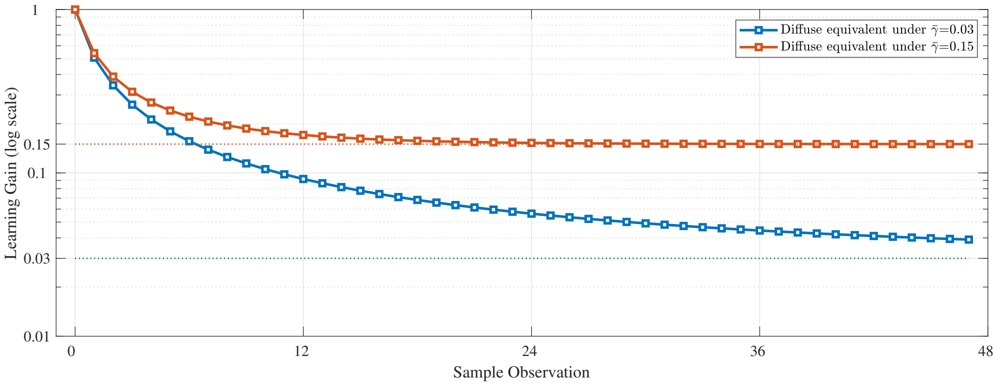
However, without a proper initialization of $\mathbf{R}_0$, the weights given to lagged observations by the CG-RLS can decay faster than the profile of weights expected from its asymptotic operation. Particularly, under diffuse initials, $\mathbf{R}_0 \to \mathbf{0}$, the CG-RLS sample weights are given by

$$\varpi_{t,l}^{dcg} = \frac{\bar{\gamma}(1-\bar{\gamma})^l}{1-(1-\bar{\gamma})^t}, \tag{8}$$

which are declining not only with the observation lag, but also with the size of the sample. These effects are illustrated in figure 1, which depicts the lagged weights given to sample observations under diffuse initials for varying sample sizes. Notice the asymptotic weights depicted in figure 1 are in fact equivalent to $\varpi_{t,l}^{cg}$. Hence, although the relative sample weights under a diffuse initial still decrease as the observation becomes outdated, the actual profile of sample weights is not time-invariant. As we will show in the next section, other than the distortion that such diffuse initials can cause to the behavioural interpretation of CG learning, such weighting distortions can generate non-negligible estimation biases in small samples.

Before turning to a quantification of such estimation biases, notice that an alternative view about the distortionary effects of diffuse initials is obtained by solving for the equivalent time-

Figure 2: Equivalent time-varying gains under diffuse initials.



Notes: The equivalent time-varying gains are calculated according to equation (9).

varying gains. Namely, equating equation (7) to equation (8) one can find that[5]

$$\tilde{\gamma}_t = \bar{\gamma}/\left(1-(1-\bar{\gamma})^t\right), \tag{9}$$

where $\tilde{\gamma}_t$ stands for the time-varying gains equivalent, in terms of information weighting, to a constant-gain $\bar{\gamma}$ under diffuse initials. The behaviour of such time-varying gains are illustrated in figure 2. Hence, the use of CG learning under diffuse initials is equivalent to the application of a decreasing sequence of gains that only converges to the underlying constant gain asymptotically.

3 Simulation Analysis

One key finding of the analysis of information weighting under least squares learning above is that the assumption of diffuse initials can distort the profile of weights given to sample observations by a constant-gain mechanism. An immediate question of interest is how much can such weighting distortions lead to biases in the estimation of models with learning. I now turn to a quantification of these effects with a simulation of the estimation of a macroeconomic model.

3.1 Model

I focus on a standard New Keynesian Phillips Curve (NKPC) model, given by

$$\begin{aligned} \pi_t &= \beta \pi^e_{t+1} + \lambda x_t + \alpha + u_t, \\ x_t &= \rho x_{t-1} + v_t, \end{aligned} \tag{10}$$

[5] See Appendix A.3 for the proof.

where π_t is inflation; π^e_{t+1} represents agents' expectations for next period inflation; x_t is a proxy for real marginal cost, usually assumed to be proportional to the labour share of income and the output gap; and, u_t is a disturbance that can be interpreted as a measurement error or as an unobserved cost-push supply shock. The forcing variable, x_t, is usually approximated empirically with measures of output gap, labour share of income, or unemployment rates (see Mavroeidis et al., 2014). The parameters in equation (10) can be interpreted as semi-structural when associated to deeper structural parameters of a micro-founded model of firms' staggered price setting (see Woodford, 2003, Ch.3). Particularly, under a Calvo framework, β stands for a discount factor, common across firms, while λ decreases with the fraction of firms that cannot update their prices in any given period, which leads to a "flatter" Phillips curve.

Under adaptive learning agents form expectations according to a PLM given by

$$\pi_t = \phi_{t-1} x_t + \varphi_{t-1} + z_t, \tag{11}$$

where ϕ_t and φ_t are parameters estimated with the RLS algorithm of (2)-(3), and are expected to converge to the rational expectations equilibrium (REE), $\phi^* = \lambda/(1-\beta\rho)$ and $\varphi^* = \alpha/(1-\beta)$, as long as $\beta < 1$ and $\beta\rho < 1$ (E-stability condition, see Evans and Honkapohja, 2001, pp. 198-200). The RE solution provides an interesting reference to simple reduced form estimates of the Phillips curve relationship between π_t and x_t. Particularly, for given α, β and λ, the implied ϕ^* and φ^* provide a description of the trade-off between inflation and, say, unemployment, after expectations have converged to equilibrium.

3.2 Simulation design

I generate 10,000 samples of artificial series of π_t and x_t assuming that $\bar{\gamma} = 0.03$, $\beta = 0.9$, $\lambda = 0.2$, $\alpha = 0$, $\rho = 0.7$, $u_t \sim N(0,3)$, and $v_t \sim N(0,1)$. For each sample I simulate the model for 2,000 observations, discarding the first 1,000 and using the remaining data for estimation of β and λ, under varying sample sizes and initial $\hat{\mathbf{R}}_0$ assumptions. Particularly, to evaluate the effect of a diffuse initial I consider estimates with: (i) the correct initial uncertainty, $\hat{\mathbf{R}}_0 = \mathbf{R}_0$, which in the artificial data is given by the estimate from observation 1,000 out of the 2,000 simulated observations; (ii) approximately diffuse initials, $\hat{\mathbf{R}}_0 = \kappa\mathbf{R}_0$, where κ is set to a small value of 10^{-4}; notice $\hat{\mathbf{R}}_0$ cannot be set exactly to zero as this would lead to degenerate estimates. All other parameters, including the learning initial ϕ_0 and the learning gain $\bar{\gamma}$, are set to their actual values – I comment on robustness checks about these assumptions below.

Since I am fixing $\bar{\gamma}$, ρ, and α, estimation of β and λ is linear on $\hat{\pi}^e_{t+1}$ and x_t, where $\hat{\pi}^e_{t+1}$ is determined by the PLM (11) and the learning estimates from the LS algorithm (2)-(3), which, in turn, depend on the simulated data, $\bar{\gamma}$, ϕ_0, and $\hat{\mathbf{R}}_0$. In order to obtain a clearer understanding of the biases introduced by the alternative $\hat{\mathbf{R}}_0$ assumptions, I conduct stepwise estimation exercises, first starting with the separate estimation of β and λ with simple regressions given by

$$(\pi_t - \lambda x_t) = \hat{\beta}\hat{\pi}^e_{t+1} + \hat{\upsilon}_{1t}, \tag{12}$$

and

$$\left(\pi_t - \beta\hat{\pi}^e_{t+1}\right) = \hat{\lambda} x_t + \hat{\upsilon}_{2t}, \tag{13}$$

respectively. On another exercise I then estimate β and λ jointly with

$$\pi_t = \hat{\beta}\hat{\pi}^e_{t+1} + \hat{\lambda} x_t + \hat{\upsilon}_{3t}. \tag{14}$$

All regressions are estimated using OLS. Christopeit and Massmann (2017) examine the asymptotic properties of the OLS estimator of structural parameters in models with learning, establishing its consistency in spite of non-standard distributions for traditional statistical inference. Interestingly, I found that the inclusion of an intercept in the estimation causes instabilities in the estimation of β – these effects can be attributed to strong persistence induced in $\hat{\pi}^e_{t+1}$ by the use of a low learning gain, as well as collinearity with the PLM intercept. Given that $\alpha = 0$ in the generated data, I estimate regressions without intercept.

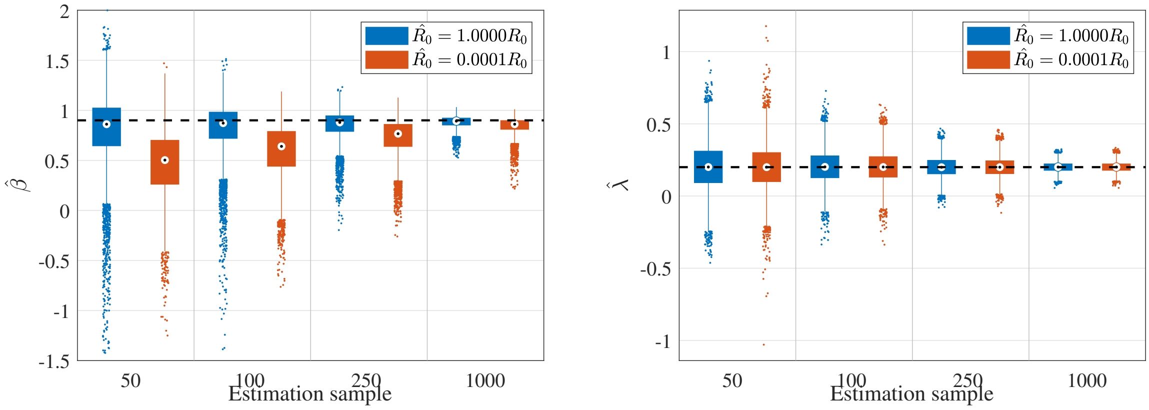
As robustness checks I also considered pre-sample estimates of ϕ_0 and $\mathbf{R}_0$ (as in Berardi and Galimberti, 2017b) and obtained similar results. I also considered the joint estimation of the learning gain, which requires non-linear estimators and is complicated by weak identification and persistent dynamics (Chevillon et al., 2010). Under constant-gain learning, RE weak identification issues are propagated as $\bar{\gamma} \to 0$ (no learning), and the collinearity between $\hat{\pi}^e_{t+1}$ and x_t increases the lower the learning gain. My experimentation with the joint estimation of $\bar{\gamma}$ indicate such estimates tend to be extremely dispersed and, under constrained estimation, concentrate on the boundaries of the pre-specified parameter space. Hence, I follow a calibration approach, fixing the learning gain so as to match survey forecasts. Particularly, $\bar{\gamma} = 0.03$ is in the range of calibrations reported by Berardi and Galimberti (2017a, Fig.8) to match survey forecasts of US CPI inflation from professionals, consumers, and policymakers.

3.3 Results

I evaluate the effect of diffuse initials by considering how alternative assumptions of $\hat{R}_0$ affect the estimates of β and λ. Starting with the individual estimation exercises, figure 3 depicts the distributions of model estimates for varying sample sizes and initials uncertainty. The β estimates are clearly downward biased under the assumption of diffuse initials, depicted in red. As will be discussed below, this finding can be directly related to the diffuse initials weighting distortions discussed in the previous section.

The distributions of the β estimates are also highly skewed towards values below the true value of the parameter, especially for small samples – that is the case even for the estimates obtained under the correct initials. This result is consistent with the findings of Chevillon et al. (2010) showing that learning generates non-standard distributions of estimates of structural

Figure 3: Individual estimates of simulated Phillips curve model with constant-gain learning.



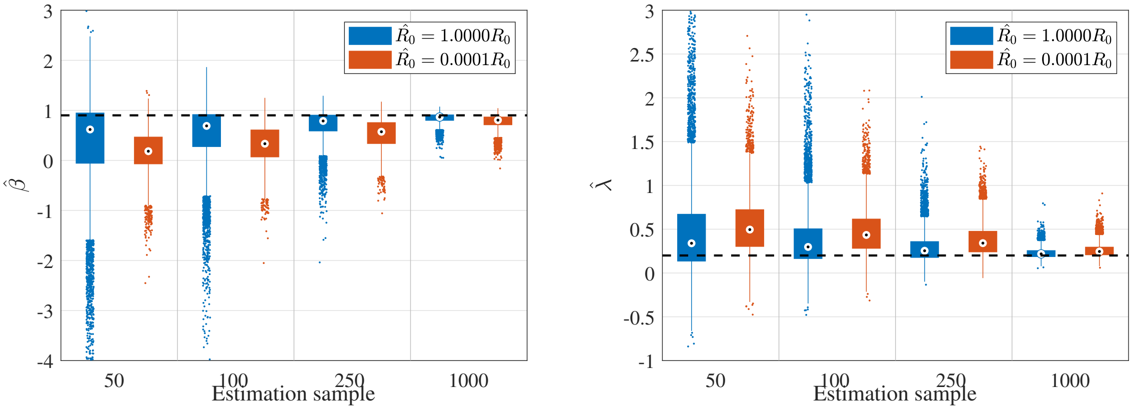
Notes: Estimates of equation (10) obtained individually, i.e., fixing λ when estimating β and vice versa, over 10,000 simulations of the model. The simulated data is generated with $\bar{\gamma} = 0.03$, $\beta = 0.9$, $\lambda = 0.2$, $\rho = 0.75$, $\sigma_u^2 = 3$, and $\sigma_v^2 = 1$. Both $\bar{\gamma}$ and ρ are prefixed to their true values during estimation. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively; the whiskers extend to ± 1.5 times the interquartile range, and estimates outside this range are considered outliers and are depicted as dots.

model parameters. Interestingly, the individual λ estimates are not affected by the varying initials, neither their distributions are affected by skewness. Hence, the inference difficulties caused by non-standard distributions under learning seem tied to the estimation of the model parameter associated with the expectations variable. As expected, these estimates tend to converge to their actual values for bigger estimation samples.

The results from the joint estimation exercise, depicted in figure 4, are similar to the previous exercise for the estimates of β, though with significant quantitative differences. For example, the median $\hat{\beta}$s estimated jointly using a diffuse initial are between 0.72 ($T = 50$) and 0.10 ($T = 1,000$) below the true value of β, while in the individual estimation exercise these medians underestimated β by 0.40 ($T = 50$) and 0.04 ($T = 1,000$). The bias also increased significantly using the correct initials, e.g., rising from -0.04 in the individual estimation exercise to -0.28 in the joint estimation one (both with $T = 50$). The estimates with the correct initials were also more strongly affected by sampling variation as reflected by the greater dispersion of estimates under the smaller estimation samples.

Another important difference in the joint estimation exercise relates to a positive bias in the estimates of λ, especially for the smaller estimation samples. The distributions of these λ estimates also show non-standard behaviour, though, in contrast to the β estimates, skewed towards values above the true value of the parameter. More importantly, here we again find that the estimates using the diffuse initials lead to greater biases. Quantitatively, the median bias in the jointly estimated $\hat{\lambda}$s using a diffuse initial (+0.30 with $T = 50$, +0.05 with $T = 1,000$)

Figure 4: Joint estimates of simulated Phillips curve model with constant-gain learning.



Notes: Same as figure 3 except that estimates are obtained jointly.

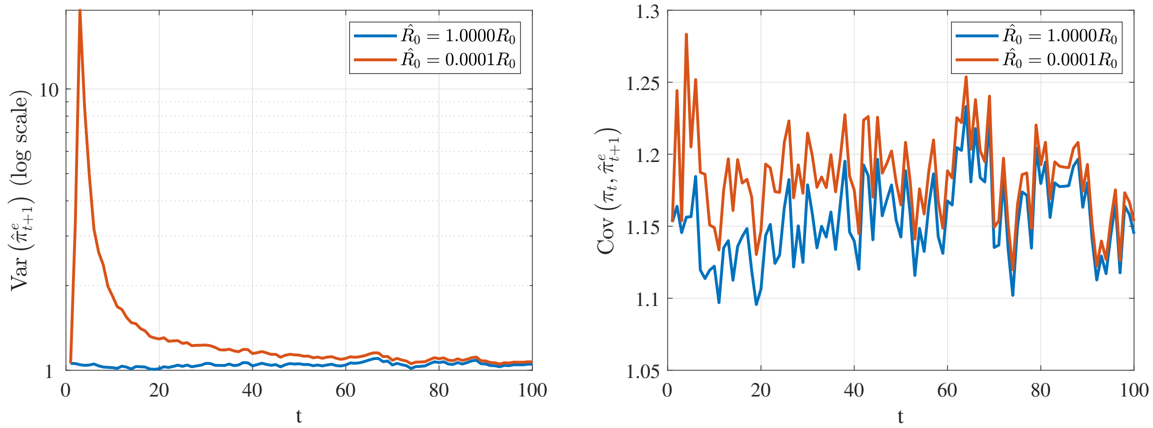
are more than twice the bias obtained with the correct initials (+0.14 with $T = 50$, +0.02 with $T = 1,000$).

Interestingly, taken together, the median biases affecting the joint estimates of β and λ tend to cancel out in the implied reduced form slope estimates of the Phillips curve under RE, $\hat{\phi}^*$. That is in contrast with the case of individual estimation of β, in the first exercise. For example, under the smaller sample, the median estimate of β in the first exercise was equal to 0.50, with diffuse initials, and 0.86, with the correct initials. Using the model RE solution, and given the fixed values of λ and ρ in that exercise, these estimates imply median $\hat{\phi}^*$'s equal to 0.32 and 0.56, respectively. Hence, the use of diffuse initials can lead to an implied Phillips curve slope almost half the magnitude (or "flatter") of its true value, $\phi^* = 0.62$. This is equivalent to saying that individual estimation of β in this model with diffuse initials, and a sample of 50 observations, can lead a researcher to predict that when the economy is producing 10% above its full-employment capacity, prices would rise at a 3.2% inflation rate, when they will in fact rise by 6.2%, plus a random shock. Hence, misspecified initials, and their associated distortions to how sample information is weighted under constant-gain learning, can lead to flatter Phillips curve estimates.

3.4 Analysis

In order to relate the biases documented above to the weighting distortions induced by a diffuse initial, it is instructive to consider the limiting behaviour of the OLS estimators of the structural model parameters. Starting with the individual estimation exercises, the OLS estimate of

Figure 5: Evolution of simulated data moments calculated across simulations.



Notes: Variances and covariances calculated for each period across the 10,000 simulations of equation (10) used in figures 3 and 4.

equation (12) is given by

$$\hat{\beta}^{ols} = \frac{Cov\,(\pi - \lambda x, \hat{\pi}^e)}{Var\,(\hat{\pi}^e)},$$
$$= \frac{Cov\,(\pi, \hat{\pi}^e) - \lambda Cov\,(x, \hat{\pi}^e)}{Var\,(\hat{\pi}^e)}, \tag{15}$$

where I drop the variables subscripts for succinctness. Clearly, $\hat{\beta}^{ols}$ would only differ, in probabilistic terms considering sampling variation, from the correct estimate of β because of deviations in $\hat{\pi}^e$ from its true value. Hence, the difference in estimation biases with respect to $\hat{\mathbf{R}}_0$ assumptions can be understood as differences caused by such assumptions in the statistical moments of the simulated data relative to the implied expectations derived from learning.

Figure 5 presents the evolution of such statistics for the simulated data – $Cov\,(x, \hat{\pi}^e)$ is not presented because it was virtually unaffected by initials assumptions. Of particular interest is the effect of increasing $Var\,(\hat{\pi}^e)$, which according to equation (15) would cause $\hat{\beta}^{ols}$ to decrease. As the LHS panel of figure 5 indicates, the use of a diffuse initial led to a substantial inflation of the variance of the implied expectations at the beginning of the estimation sample, which explains the downward bias observed in the estimates of β under diffuse initials. This finding is also consistent with the analysis of the previous section showing that the diffuse initials lead to an overweighting of initial sample observations, or, equivalently, to an increase in the initial learning gains. As is well known, a higher learning gain leads to more volatile learning estimates (see, e.g., Evans and Honkapohja, 2001). Hence, the use of diffuse initials lead to more volatile learning estimates and their implied expectations, which ultimately translates into more biased estimates of the relevance of expectations in this model.

Similar analysis can be applied to the other model estimates. For the second exercise, the OLS estimate of equation (13) is given by

$$\begin{aligned}\hat{\lambda}^{ols} &= \frac{Cov\left(\pi - \beta\hat{\pi}^{e}, x\right)}{Var\left(x\right)} \\ &= \frac{\overline{Cov\left(\pi, x\right)} - \beta Cov\left(x, \hat{\pi}^{e}\right)}{\overline{Var\left(x\right)}}.\end{aligned} \tag{16}$$

In contrast to the first exercise, the individual OLS estimate of λ is not affected by the variance of the expectations variable – to facilitate analysis, terms not affected by the initials are depicted with a upper bar. Here, the only component that may cause differences between the initial assumptions is the covariance between the exogenous variable and the implied expectations, $Cov\left(x, \hat{\pi}^{e}\right)$. As the RHS panel of figure 5 indicates, this statistic was not strongly affected by the use of a diffuse initial, which explains why there was no significant difference observed in the estimates reported for this exercise, in the RHS panel of figure 3.

Finally, for the joint estimation exercise the corresponding OLS estimates are given by

$$\hat{\beta}^{ols} = \frac{\overline{Var\left(x\right)}Cov\left(\pi, \hat{\pi}^{e}\right) - Cov\left(x, \hat{\pi}^{e}\right)\overline{Cov\left(\pi, x\right)}}{Var\left(\hat{\pi}^{e}\right)\left(\overline{Var\left(x\right)} - Cov\left(x, \hat{\pi}^{e}\right)\right)}, \tag{17}$$

$$\hat{\lambda}^{ols} = \frac{Var\left(\hat{\pi}^{e}\right)\overline{Cov\left(\pi, x\right)} - Cov\left(x, \hat{\pi}^{e}\right)Cov\left(\pi, \hat{\pi}^{e}\right)}{\overline{Var\left(x\right)}\left(Var\left(\hat{\pi}^{e}\right) - Cov\left(x, \hat{\pi}^{e}\right)\right)}. \tag{18}$$

Although in this case the effects become more convoluted, it is clear that: (i) the variance of expectations still has a negative effect on the β estimates if $Var\left(x\right) > Cov\left(x, \hat{\pi}^{e}\right)$, which was the case in the model simulation presented here; (ii) the estimates of λ are now also affected by initials uncertainty through its effects on the variance of expectations; particularly, it can be shown that when $Cov\left(x, \hat{\pi}^{e}\right) > 0$ (generally true for model 10 given that x enters the PLM) and $Cov\left(\pi, \hat{\pi}^{e}\right) > Cov\left(\pi, x\right)$ (also generally the case for $\beta > \lambda$), the λ estimates will be positively affected by the increasing variance of expectations associated with the diffuse initials. These two points offer an explanation for the biases caused by the diffuse initials reported in figure 4.

4 Empirical Application

I now turn to an empirical evaluation of the effects of diffuse initials on the estimation of the standard NKPC model with constant-gain learning. As in the simulation analysis of the previous section, the main focus of this empirical exercise is on the effect that initials uncertainty can have on the estimates of the model parameters. Particularly, I again consider two alternatives for the initial matrix of second moments: (i) an estimate obtained with the pre-sample data to represent the "correct" initials, henceforth denoted as the non-diffuse initials – details about this initialization are provided below; and, (ii) a downward re-scaled version of (i) to represent the diffuse initials approach.

4.1 Data and estimation approach

I use U.S. quarterly data covering the period from 1947 to 2019, focusing on estimates of β and λ across decade sub-samples. The focus on sub-samples allows an analysis of the stability of the Phillips curve relationship, which has historically attracted great interest in the literature (see, e.g., Gordon, 2011). Another known issue with empirical estimates of the NKPC relates to their sensitivity with respect to the data definitions of the measures of price inflation and production slack (see Mavroeidis et al., 2014).

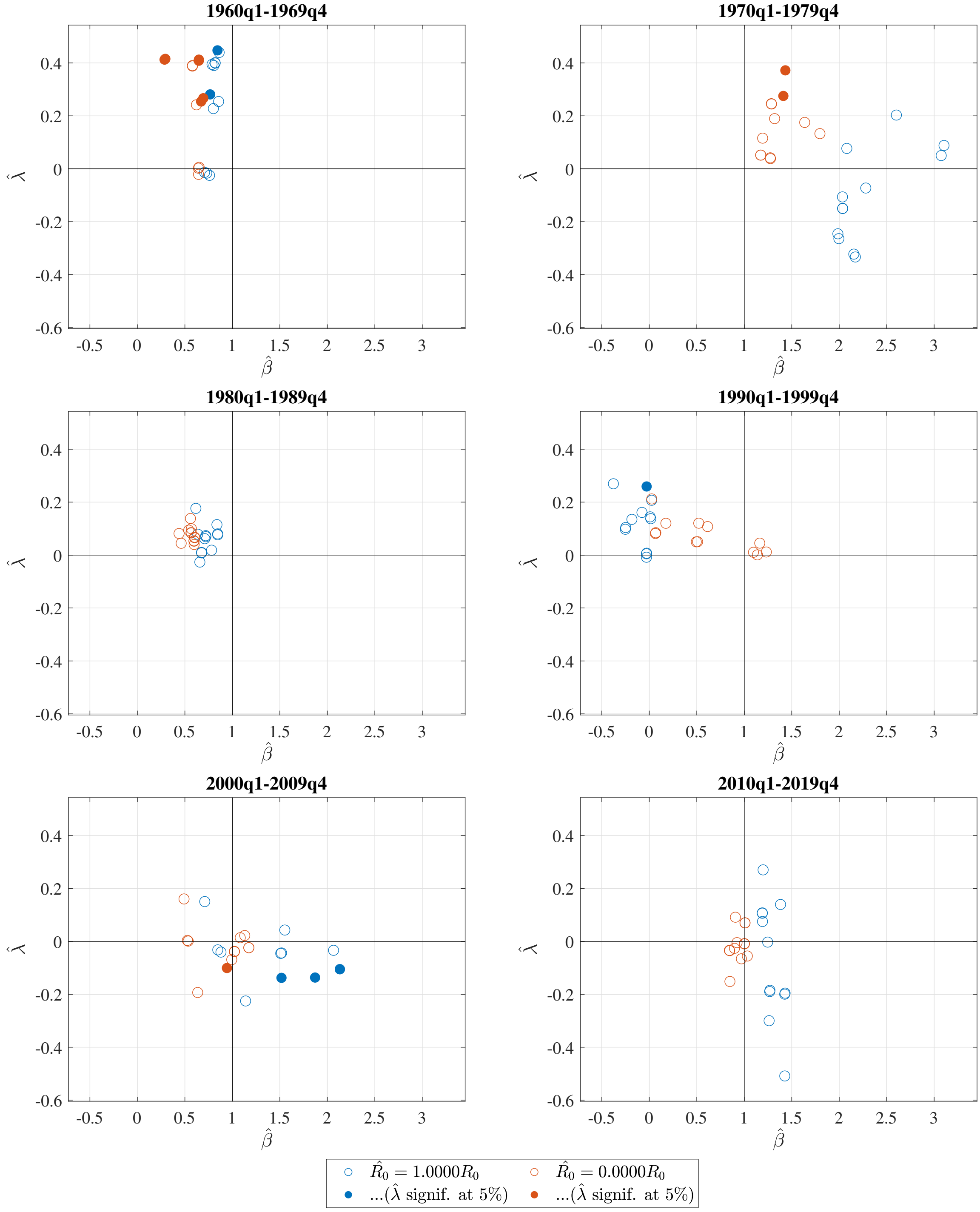
To deal with such specification uncertainty, I consider combinations of three inflation measures, based on the CPI, the core CPI, and the GDP deflator, with four alternative proxies for real marginal costs, namely, an output gap measure based on real GDP data, non-farm business sector labour shares, unemployment rates, and the unemployment rate gap relative to an estimate of the natural rate. Inflation rates are annualised by multiplying the quarterly rates by four. For comparative purposes, all measures of x_t are filtered using the Hodrick-Prescott filter (with $\lambda^{HP} = 1,600$), and all measures, including the inflation rates, are standardized prior to estimation to have zero mean and variance equal to unity. All data series are obtained from the FRED database of the St. Louis Fed.

To be consistent with the previous simulation analysis, all model estimates are obtained using OLS. I focus on the joint estimation of β and λ, estimating regressions of the form of equation (14), while pre-fixing the other parameters to plausible values: $\bar{\gamma} = 0.03$ is again fixed according to the calibrations reported by Berardi and Galimberti (2017a) to match survey forecasts; ρ is pre-estimated by fitting a first-order autoregression on the full-sample of each measure used as x_t; the learning initials $\{\phi_0, \varphi_0, \mathbf{R}_0\}$ are estimated over pre-sample data using WLS in order to obtain initials consistent with the constant-gain learning adopted in the estimation sample (see Berardi and Galimberti, 2017b).

4.2 Results and analysis

Figures 6 and 7 present the estimation results. There is substantial variation in the model estimates across the variables definitions and the sub-samples. The λ estimates are mostly consistent with their expected signs up to the end of the 20th century, although rarely with statistical significance (depicted with a filled marker). In contrast, the majority of the sub-sample β estimates are statistically significant at the 5% significance level (not depicted). However, as discussed above, such inferences should be interpreted with caution considering that learning can generate non-standard distributions of statistical tests (see also Chevillon et al., 2010; Christopeit and Massmann, 2017).

Figure 6: Estimates of U.S. Phillips curve with constant-gain learning by decades.



Notes: Estimates of equation (14) obtained using different combinations of data definitions for inflation, $\pi_t = \{\text{CPI, core CPI, GDP deflator}\}$, and proxy for real marginal cost, $x_t = \{\text{real GDP gap, labour share, unemployment, natural rate of unemployment gap}\}$. For comparative purposes, both π_t and x_t are standardized prior to estimation, and estimates with unemployment as x_t are depicted as $-\hat{\lambda}$. All estimates obtained under a fixed learning gain, $\bar{\gamma} = 0.03$. Statistical significance at the 5% level are depicted for $\hat{\lambda}$ with filled markers and are based on HAC standard errors. Such inferences under learning should be interpreted with caution since estimators distributions can become nonstandard.

According to the model estimates, the 1960s may be considered as the "golden days" of the Phillips curve, as several λ estimates display statistical significance and signs according to expectations. Similarly, most β estimates in the 1960s sub-sample are statistically significant and below unity, hence satisfying E-stability conditions in this model. The estimates for the 1970s, in contrast, indicate an important change on the estimates associated with the forward-looking expectations in this model. Namely, the β estimates increase above unity, also with a more robust increase using the non-diffuse initials, which suggests a period of unstable inflation expectations relative to observed inflation rates. The β estimates then return to the E-stability range during the 1980s, while the λ estimates become less dispersed between the values of 0 and 0.2.

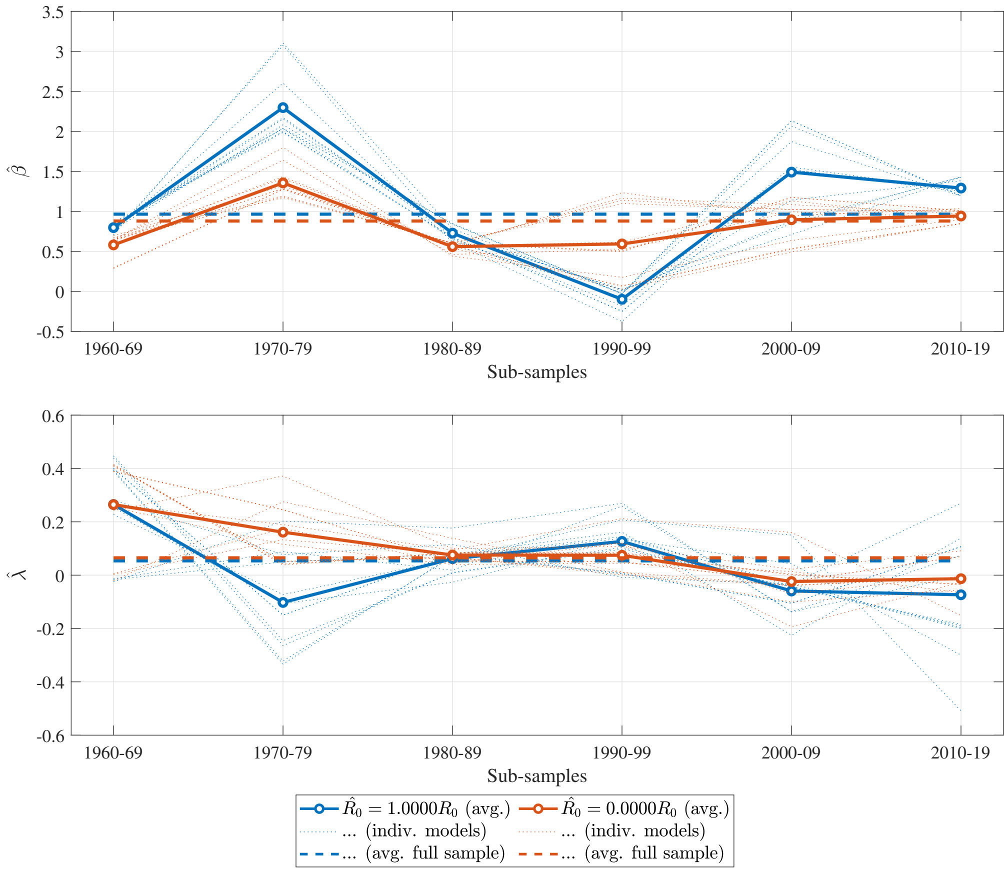
These results are consistent with the view that a strong correlation between inflation and economic activity may have misled policymakers to believe on an apparent trade-off between inflation and unemployment in the 1960s. The associated decline of active stabilization policies then led to an increase in inflation expectations in the 1970s, here reflected as a period of unstable expectations, which ultimately increased actual inflation. This is the so-called Great Inflation period, which prompted the monetary authority to revert to a more active policy of inflation and expectations stabilization in the 1980s (see, e.g., Orphanides and Williams, 2005; Primiceri, 2006; Sargent et al., 2006).

The estimates from the 1990s reflect a period of decreased relevance of inflation expectations, a result that, again, seems more robust with the use of the non-diffuse initials. This result is indicative of a build-up of credibility in the monetary authority resolve to keep inflation stable. Interestingly, in the 2000s and 2010s, the β estimates jump again outside the E-stability range, especially for the estimates based on non-diffuse initials. At the same time, the λ estimates become more dispersed and move towards negative values. This is consistent with previous evidence in the learning literature that decreasing beliefs about inflation persistence provide an explanation for lower average and volatility of inflation during the so-called Great Moderation period (1986–2006) in the U.S, as well as the flattening of the Phillips curve (Slobodyan and Wouters, 2012).

More important to the purposes of this paper, the empirical estimates of the NKPC are found to depend on the assumption about the initials uncertainty. The evolution of the averages across the the different specifications, presented in figure 7, indicate that the β estimates tend to be less sensitive to the sub-samples under the diffuse initials, hence less informative about violations of expectational stability over time. For the λ estimates, the use of diffuse initials also point to a smoother flattening of the Phillips curve relative to the estimates with non-diffuse initials.

Table 1 presents another comparative between these estimates, focusing on one particular specification that uses the GDP deflator for inflation and the natural rate of unemployment gap as a proxy for real marginal costs – hence the expected sign of λ is negative. The β estimates obtained under the diffuse initials are mostly smaller than those obtained with the non-diffuse

Figure 7: Evolution of estimates of U.S. Phillips curve with constant-gain learning by decades.



Notes: Same as figure 6.

Table 1: Empirical estimates of a U.S. Phillips curve with constant-gain learning by decades.

	1960-19 (240 qtrs.)	1960-69 (40 qtrs.)	1970-79 (40 qtrs.)	1980-89 (40 qtrs.)	1990-99 (40 qtrs.)	2000-09 (40 qtrs.)	2010-19 (40 qtrs.)
- Under pre-initialized $\hat{R}_0$:							
$\hat{\beta}$	0.973	0.843	1.986	0.676	-0.252	1.518	1.270
	(0.207)	(0.207)	(0.295)	(0.314)	(0.384)	(0.301)	(0.165)
$\hat{\lambda}$	-0.031	-0.447	0.246	-0.008	-0.097	0.044	0.185
	(0.094)	(0.214)	(0.225)	(0.088)	(0.156)	(0.063)	(0.107)
$\hat{\phi}$	-0.241	-1.823	-0.316	-0.021	-0.080	-0.122	-1.354
	(0.762)	(0.998)	(0.265)	(0.229)	(0.139)	(0.180)	(1.144)
- Under diffuse $\hat{R}_0$:							
$\hat{\beta}$	0.893	0.646	1.173	0.603	1.097	1.022	1.001
	(0.174)	(0.096)	(0.154)	(0.179)	(0.350)	(0.183)	(0.127)
$\hat{\lambda}$	-0.036	-0.408	-0.052	-0.066	-0.010	0.038	0.008
	(0.090)	(0.153)	(0.134)	(0.087)	(0.094)	(0.072)	(0.083)
$\hat{\phi}$	-0.179	-0.968	1.047	-0.143	-0.565	0.440	0.080
	(0.457)	(0.326)	(4.857)	(0.221)	(6.851)	(1.239)	(0.858)

Notes: Estimates of equation (14) obtained using GDP deflator for inflation and the natural rate of unemployment gap as a proxy for real marginal cost. Both variables are standardized prior to estimation. All estimates obtained under a fixed learning gain, $\bar{\gamma} = 0.03$. The implied reduced form slope of the Phillips curve, $\hat{\phi}$, is obtained according to the REE. Standard errors in parentheses are HAC robust. Inference under learning should be interpreted with caution since estimators distributions can become non-standard.

ones, except for the 1990-99 decade, when inflation expectations are found to lose significance. The λ estimates obtained with the two alternative initials assumptions move in different directions throughout the sub-samples, but agree on their sign and statistical significance for the 1960s. In the last two decades, $\hat{\lambda}$ turns positive under both initials, but with a greater increase and statistical significance (at the 10% level) under the non-diffuse initials for the 2010s. Nevertheless, in conjunction with the results for $\hat{\beta}$, the implied reduced form slope of the Phillips curve is always negative under the non-diffuse initials, consistent with expectations about this relationship, while the estimates under diffuse initials imply inverted Phillips curves during the 1970s, 2000s and 2010s.

5 Concluding remarks

In this paper I proposed a more general non-recursive representation of the recursive least squares algorithm that is used in the adaptive learning literature to represent how agents form their expectations in economic settings. According to this new formulation, the recursive learning mechanism is more properly represented by a penalized weighted least squares estimator, where a penalty term accounts for the effects of the learning initial estimates. The non-recursive formulation also allowed a renewed analysis of how information is weighted in the implied estimates of agents' perceived law of motion. Such weights are directly determined by the sequence of learning gains used in the recursive least squares algorithm, and the specification of the uncertainty around initial learning estimates. The framework proposed in this paper provides flexible analytical expressions for the calculation of information weighting under different assumptions on the evolution of the learning gains and initial beliefs.

One important finding obtained under this refreshed framework is that, without a proper account for the learning initial, the estimation of models under the assumption of a constant gain over increasing samples of data would imply agents give a decreasing weight to more recent observations, distorting the real-time tracking interpretation of this mechanism. The relevance of this distortion was evidenced by simulation and empirical exercises, where the misspecified initials led to a systematic bias to estimates on the relevance of expectations in a Phillips curve model. These biases also affected estimates of the responsiveness of inflation rates to output gaps. Hence, a proper account of how information is weighted under alternative learning mechanisms and assumptions about initial beliefs are important aspects for the estimation of models of imperfect information such as adaptive learning.

References

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 (2017a). Empirical calibration of adaptive learning. *Journal of Economic Behavior & Organization 144*, 219 – 237.

Berardi, M. and J. K. Galimberti (2017b). On the initialization of adaptive learning in macroeconomic models. *Journal of Economic Dynamics and Control 78*, 26 – 53.

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

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

Christopeit, N. and M. Massmann (2017). Estimating structural parameters in regression models with adaptive learning. *Econometric Theory*, 1–44.

Evans, G. W. and S. Honkapohja (1993). Adaptive forecasts, hysteresis, and endogenous fluctuations. *Economic Review*, 3–13.

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

Evans, G. W., S. Honkapohja, and N. Williams (2010). Generalized stochastic gradient learning. *International Economic Review 51*(1), 237–262.

Galimberti, J. K. (2019). An approximation of the distribution of learning estimates in macroeconomic models. *Journal of Economic Dynamics & Control 102*, 29–43.

Gordon, R. J. (2011). The history of the phillips curve: Consensus and bifurcation. *Economica 78*(309), 10–50.

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

Lubik, T. A. and C. Matthes (2016). Indeterminacy and learning: An analysis of monetary policy in the great inflation. *Journal of Monetary Economics 82*, 85 – 106.

Malmendier, U. and S. Nagel (2016). Learning from inflation experiences. *Quarterly Journal of Economics 131*(1), 53–87.

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

Marcet, A. and T. J. Sargent (1989). Convergence of least squares learning mechanisms in self-referential linear stochastic models. *Journal of Economic Theory 48*(2), 337–368.

Markiewicz, A. and A. Pick (2014). Adaptive learning and survey data. *Journal of Economic Behavior & Organization 107, Part B*(0), 685 – 707. Empirical Behavioral Finance.

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

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

Primiceri, G. E. (2006). Why inflation rose and fell: Policy-makers' beliefs and u. s. postwar stabilization policy. *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 (2012). Learning in a medium-scale dsge model with expectations based on small forecasting models. *American Economic Journal: Macroeconomics 4*(2), 65–101.

Williams, N. (2019). Escape dynamics in learning models. *Review of Economic Studies 86*(2), 882–912.

Woodford, M. (2003). *Interest and Prices*. Princetion, NJ: Princeton University Press.

A Proofs and Derivations

A.1 Correspondence between penalized WLS and RLS

To see how the RLS of (2)-(3) can be derived from the penalized WLS formulation of (5) and (6), first notice that iterating (3) recursively from $\mathbf{R}_0$ we have that

$$\mathbf{R}_t = \sum_{i=1}^{t} \omega_{t,i}\mathbf{x}_i\mathbf{x}_i' + \omega_{t,0}\mathbf{R}_0,$$

which is the inverse of the first term in (5), leading to

$$\hat{\phi}_t = \mathbf{R}_t^{-1}\left[\sum_{i=1}^{t} \omega_{t,i}\mathbf{x}_i y_i + \omega_{t,0}\mathbf{R}_0\phi_0\right]. \tag{19}$$

For the second term notice that

$$\begin{aligned}\sum_{i=1}^{t} \omega_{t,i}\mathbf{x}_i y_i &= \sum_{i=1}^{t-1} \omega_{t,i}\mathbf{x}_i y_i + \gamma_t\mathbf{x}_t y_t,\\ &= (1-\gamma_t)\sum_{i=1}^{t-1} \omega_{t-1,i}\mathbf{x}_i y_i + \gamma_t\mathbf{x}_t y_t,\end{aligned}$$

and

$$\omega_{t,0}\mathbf{R}_0\phi_0 = (1-\gamma_t)\,\omega_{t-1,0}\mathbf{R}_0\phi_0,$$

where we use

$$\omega_{t,i} = (1-\gamma_t)\,\omega_{t-1,i},$$

which follows from (6). Hence, (19) is equivalent to

$$\hat{\phi}_t = \mathbf{R}_t^{-1}\left[\gamma_t\mathbf{x}_t y_t + (1-\gamma_t)\left(\sum_{i=1}^{t-1} \omega_{t-1,i}\mathbf{x}_i y_i + \omega_{t-1,0}\mathbf{R}_0\phi_0\right)\right]. \tag{20}$$

Lagging (19) one period we find that

$$\mathbf{R}_{t-1}\hat{\phi}_{t-1} = \sum_{i=1}^{t-1} \omega_{t-1,i}\mathbf{x}_i y_i + \omega_{t-1,0}\mathbf{R}_0\phi_0,$$

which can be substituted into (20) to yield

$$\hat{\phi}_t = \mathbf{R}_t^{-1}\left[\gamma_t \mathbf{x}_t y_t + (1-\gamma_t)\,\mathbf{R}_{t-1}\hat{\phi}_{t-1}\right]. \tag{21}$$

From (3) notice that

$$(1-\gamma_t)\,\mathbf{R}_{t-1} = \mathbf{R}_t - \gamma_t \mathbf{x}_t \mathbf{x}_t',$$

which substituted into (21) and after rearranging leads to

$$\begin{aligned}
\hat{\phi}_t &= \mathbf{R}_t^{-1}\left[\gamma_t \mathbf{x}_t y_t + \left(\mathbf{R}_t - \gamma_t \mathbf{x}_t \mathbf{x}_t'\right)\hat{\phi}_{t-1}\right], \\
&= \gamma_t \mathbf{R}_t^{-1}\mathbf{x}_t y_t + \hat{\phi}_{t-1} - \gamma_t \mathbf{R}_t^{-1}\mathbf{x}_t \mathbf{x}_t' \hat{\phi}_{t-1}, \\
&= \hat{\phi}_{t-1} + \gamma_t \mathbf{R}_t^{-1}\mathbf{x}_t\left(y_t - \mathbf{x}_t'\hat{\phi}_{t-1}\right),
\end{aligned}$$

establishing the correspondence between the penalized WLS solution of (5) and the RLS of (2).

A.2 Absolute and relative weights

Letting W_t^n stand for the sum of weights starting from weight n up to weight t, from the definition of the absolute weights, (6), this sum of weights can be expanded according to

$$\begin{aligned}
W_t^0 &= \sum_{i=0}^{t} \omega_{t,i}, \\
&= \prod_{j=1}^{t}(1-\gamma_j) + \sum_{i=1}^{t-1}\gamma_i \prod_{j=i+1}^{t}(1-\gamma_j) + \gamma_t.
\end{aligned} \tag{22}$$

Expanding the first term of (22) we have that

$$\begin{aligned}
\omega_{t,0} &= (1-\gamma_1)(1-\gamma_2)\dots(1-\gamma_{t-1})(1-\gamma_t), \\
&= (1-\gamma_2)\dots(1-\gamma_{t-1})(1-\gamma_t) - \gamma_1\prod_{j=2}^{t}(1-\gamma_j), \\
&= 1 - \gamma_t - \sum_{i=1}^{t-1}\gamma_i \prod_{j=i+1}^{t}(1-\gamma_j).
\end{aligned} \tag{23}$$

Returning to (22) we then have

$$
\begin{aligned}
W_t^0 &= 1-\gamma_t-\sum_{i=1}^{t-1}\gamma_i\prod_{j=i+1}^{t}(1-\gamma_j)+\sum_{i=1}^{t-1}\gamma_i\prod_{j=i+1}^{t}(1-\gamma_j)+\gamma_t,\\
&= 1.
\end{aligned}
$$

A.3 Equivalent time-varying gains under diffuse initials

The sequence of gains, $\tilde{\gamma}_t$, that generates equivalent weightings as a constant-gain under diffuse initials needs to solve

$$
\begin{aligned}
\varpi_{t,l} &= \varpi_{t,l}^{dcg},\\
&= \frac{\bar{\gamma}(1-\bar{\gamma})^l}{1-(1-\bar{\gamma})^t}
\end{aligned}
\tag{24}
$$

for all t and l. From equation (7), starting with $l=0$ we simply have that

$$
\tilde{\gamma}_t = \bar{\gamma}/\left(1-(1-\bar{\gamma})^t\right). \tag{25}
$$

It only remains to validate if equation (25) also solves equation (24) for $l>0$. Substituting equation (25) into equation (7) for $0<l<t$,

$$
\begin{aligned}
\varpi_{t,l} &= \frac{\bar{\gamma}}{1-(1-\bar{\gamma})^{t-l}}\prod_{j=0}^{l-1}\left(1-\frac{\bar{\gamma}}{1-(1-\bar{\gamma})^{t-j}}\right),\\
&= \frac{\bar{\gamma}(1-\bar{\gamma})^l}{1-(1-\bar{\gamma})^{t-l}}\prod_{j=0}^{l-1}\left(\frac{1-(1-\bar{\gamma})^{t-j-1}}{1-(1-\bar{\gamma})^{t-j}}\right),\\
&= \frac{\bar{\gamma}(1-\bar{\gamma})^l}{1-(1-\bar{\gamma})^{t-l}}\left(\frac{1-(1-\bar{\gamma})^{t-l}}{1-(1-\bar{\gamma})^{t}}\right),\\
&= \frac{\bar{\gamma}(1-\bar{\gamma})^l}{1-(1-\bar{\gamma})^{t}},
\end{aligned}
$$

which solves equation (24) for $l>0$. Finally, notice that under a diffuse initial the weight given to the learning initial is null, i.e., $\varpi_{t,t}^{dcg}=0$. This is equivalent to using a $\gamma_1=1$, which is again satisfied by equation (25).