Discussion Paper Series

On the plausibility of adaptive learning in macroeconomics: A puzzling conflict in the choice of the representative algorithm

By

Michele Berardi and Jaqueson K. Galimberti

Centre for Growth and Business Cycle Research, Economic Studies,
University of Manchester, Manchester, M13 9PL, UK

November 2012
Number 177

On the plausibility of adaptive learning in macroeconomics: A puzzling conflict in the choice of the representative algorithm[*]

Michele Berardi

The University of Manchester

Jaqueson K. Galimberti[†]

The University of Manchester and The Capes Foundation

November 7, 2012

Abstract

The literature on bounded rationality and learning in macroeconomics has often used recursive algorithms such as least squares and stochastic gradient to depict the evolution of agents' beliefs over time. In this work, we try to assess the plausibility of such practice from an empirical perspective, by comparing forecasts obtained from these algorithms with survey data. In particular, we show that the relative performance of the two algorithms in terms of forecast errors depends on the variable being forecasted, and we argue that rational agents would therefore use different algorithms when forecasting different variables. By using survey data, then, we show that agents instead always behave as least squares learners, irrespective of the variable being forecasted. We thus conclude that such findings point to a puzzling conflict between rational and actual behaviour when it comes to expectations formation.

Keywords: expectations, learning algorithms, forecasting, least squares, stochastic gradient.

JEL codes: C63, D84, E37.

1 Introduction

Adaptive learning algorithms have been proposed to provide an alternative to, and a justification for, rational expectations (RE) equilibria in macroeconomics. Going beyond the RE hypothesis, however, comes at the cost of introducing another degree of freedom in macroeconomic modeling, since one has to be specific about which algorithm is assumed to represent agents' behavior.

[*] We thank, without implicating, for useful comments received: James Banks, Keith Blackburn, George Bratsiotis, Katsushi Imai, Kyriakos Neanidis, and in particular Anne Villamil, who co-supervises the doctoral studies of Jaqueson K. Galimberti.

[†] Corresponding author. E-mail: jaqueson.galimberti@postgrad.manchester.ac.uk.

The usual choice for this purpose has been the least squares (LS) algorithm, possibly due to its widespread popularity between econometricians. Apart from this vogue, though, there seems to be no clear justification for such choice and, taking a bounded rationality standpoint, one could for example argue for computationally simpler alternatives such as the stochastic gradient (SG) algorithm, which has also received some attention in the literature.

Since the ultimate sieve of scientific research comes with empirical validation, we argue that the previous literature has neglected the need of a realistic justification in the choice of a learning algorithm. Our main contribution is therefore an attempt to fill that gap, and we do this by proposing an empirical assessment on the relative plausibility of the LS and SG algorithms as representative of agents' behavior.

In the choice of an algorithm for the purpose of representing agents' learning behaviour, two alternative approaches are possible: one that would select the algorithm that performs better, therefore emphasizing the rational aspect of agents' choice, and the other that would instead select the algorithm that better resembles actual forecasts, in an attempt to mimic agents' actual behaviour. It is not clear whether these two alternative approaches are compatible with each other.

In order to shed some light on this point, we propose a framework that allows to investigate empirically the plausibility of each algorithm along the two dimensions discussed. Using such framework, together with US macroeconomic data, we then try to answer our main question regarding the plausibility of different learning algorithms for the purpose of representing agents' forecasting behaviour. Our findings reveal a puzzle, since agents' actual behavior, as observed from surveys data, appears to be in conflict with what would have been done by agents wanting to optimise their forecasting performance.

1.1 Approach and preview of results

In order to compare the relative plausibility of the LS and the SG algorithms as representative of agents' learning mechanism, we analyse the quality of their forecasts along two dimensions, in two related exercises.

From the agents' perspective, one could argue that what matters is the accuracy of the forecasts delivered by each algorithm when compared to actual observations, and therefore the first exercise is an evaluation of the algorithms' *forecasting performance*. From the researcher's perspective, instead, what matters is arguably the closeness of the algorithms' forecasts to the observed behavior of economic agents, as represented by surveys data, and therefore our second exercise evaluates the algorithms' *resemblance to surveys*.

We carry out these exercises using real-time quarterly data on US inflation and output growth covering a broad post-WWII period of time, from 1947q2 to 2011q4. Our first exercise indicates that the SG algorithm is to be preferred when forecasting inflation, while the LS has a superior forecasting performance for growth, thus showing that the statistical properties of the data series under scrutiny do matter for the suitability of different algorithms.

In the second exercise, however, our results indicate an intriguing dominance for the LS algorithm, as this is found to provide forecasts closer to those observed from surveys for both inflation and growth variables. Taken alone, such a result is maybe not surprising, and it actually goes in support of the related literature we discuss below. The puzzling aspect of this result, though, comes from its joint interpretation with the previous finding regarding the performance comparison. Namely, while the rational choice of an algorithm is found to depend on the variable of interest, agents are found to behave as LS learners irrespective of the variable forecasted.

1.2 Related literature

This paper is related to the literature on learning and expectations in macroeconomics (Evans and Honkapohja, 2001). Our contribution is mainly relevant for the applied[1] side of this literature, where emphasis has been given to the role of learning in explaining macroeconomic fluctuations. One of the seminal contributions in this field was given by Sargent (1999), and a (non-exhaustive) list of recent works includes Huang et al. (2009); Eusepi and Preston (2011); Milani (2011).

The key feature in these works is represented by the replacement of assumptions implying an instantaneous adjustment of agents' expectations with a characterization of agents as adaptive learners. Also common to most of them is the use of the LS algorithm for such purpose[2], and it is at this point that our study becomes relevant: by inquiring about the choice of a representative algorithm, our study brings up evidence of a conflict between rationality assumptions and the actual behavior of agents in forming expectations.

Our work is also relevant at a theoretical level for the recent literature advocating for the use of learnability conditions as a solution for RE indeterminacy (see, e.g., McCallum, 2007; Bullard and Eusepi, 2009; Ellison and Pearlman, 2011). One key aspect of these studies is the use of the so-called E-stability[3] conditions for the purpose of solving RE indeterminacy: drawing upon a tight link between E-stability and LS-learnability, the learning explanation has been adopted as the foundation for an equilibrium selection criterion.

The main problem with this argument is in its reliance on the link between E-stability and LS-learnability, as it is now understood that the learnability of a given RE equilibrium is defined with respect to the algorithm of choice. Analyses of learning using the SG algorithm, for instance, pointed to convergence conditions distinct to those obtained under LS learning (Barucci and Landi, 1997; Heinemann, 2000; Giannitsarou, 2005). An assessment of the plausibility of these different learning algorithms

[1] Here applied is taken to encompass both simulations and exercises of empirical estimation or calibration.

[2] There were some few exceptions, as in Marcet and Nicolini (2003) and Bullard and Eusepi (2005).

[3] "A RE equilibrium is said to be expectationally stable if, given a small deviation of expectations functions from rationality, the system returns to the equilibrium under a natural revision rule." (Evans, 1985, p. 1218)

is therefore of central relevance to the theoretical debate over learnability as a solution to RE indeterminacy.

Finally, our work is also linked to the literature investigating how different learning rules perform in forecasting macroeconomic data and in matching forecasts from surveys. Such literature includes the works of Branch and Evans (2006) and Weber (2010), which provided applications to US and European data, respectively. Although adopting an approach similar to ours, these works did not provide a comparison between the LS and SG algorithms, but mainly focused on the LS and its different calibrations.

2 Computational Framework

2.1 Preliminaries

Consider the estimation context faced by a real-time agent wishing to obtain inferences about the law of motion of a variable of interest. From an economic perspective, these inferences can be thought of as the middle step agents undertake in a process of learning-to-forecast in order to form their expectations. Our focus in this paper is on the comparison of the forecasts obtained from linear models estimated by the LS and the SG algorithms. It is the purpose of this section to define the estimation framework under which these algorithms are assumed to operate.

We use a common framework for the computational implementation of the two learning algorithms. Following a state-space approach, we combine a random walk hypermodel with an unrestricted vector autoregression (VAR) model. This latter specifies how the macro variables are (dynamically) related, in a non-structural sense, whereas the random walk is taken as representative of the time-varying structure of the VAR coefficients (see, e.g., Stock and Watson, 1996). Under Gaussianity assumptions, optimal Kalman filtering results can be readily associated to this state-space representation (see, e.g., Ljung and Soderstrom, 1983).

In order to unify our computational approach, we then adopt the exact correspondences between the learning algorithms and the Kalman recursions recently established by Berardi and Galimberti (2013). These correspondences turn out to be useful for our understanding of how the algorithms relate to each other in a statistical sense, and such an understanding will be crucial for interpreting the results we obtain later in comparing their associated forecasts.

2.2 State-space unifying representation

Let $\mathbf{y}_t = (y_{1,t}, \ldots, y_{N,t})'$ and $\mathbf{x}_t = (x_{1,t}, \ldots, x_{K,t})'$ denote vectors of period t observations on the N and K endogenous and explanatory variables of interest, respectively. Then let $\mathbf{y}_t$ and $\mathbf{x}_t$ be related through the law of motion given by the system of equations

$$y_{j,t} = \mathbf{x}_t' \boldsymbol{\theta}_{j,t} + \varepsilon_{j,t},\ \forall j = 1, \ldots, N, \tag{2.1}$$

where $\boldsymbol{\theta}_{j,t}$ are vectors of time-varying parameters containing the coefficients that relate the endogenous variable $y_{j,t}$ to the K explanatory variables $\mathbf{x}_t$.

The tracking requirements posed by the above time-varying environment motivates the need for a recursive solution. Two of the main forms adopted in the literature to represent agents' learning-to-forecast behavior are the LS and the SG algorithms.

Algorithm 1 (LS)**.** *Under the estimation context of* (2.1)*, the* LS *algorithm assumes the form of*[4]

$$\begin{aligned} \hat{\boldsymbol{\theta}}_t^{LS} &= \hat{\boldsymbol{\theta}}_{t-1}^{LS} + \gamma_t \mathbf{R}_t^{-1} \mathbf{x}_t \left(y_t - \mathbf{x}_t' \hat{\boldsymbol{\theta}}_{t-1}^{LS} \right), & (2.2) \\ \mathbf{R}_t &= \mathbf{R}_{t-1} + \gamma_t \left(\mathbf{x}_t \mathbf{x}_t' - \mathbf{R}_{t-1} \right), & (2.3) \end{aligned}$$

where γ_t is a learning gain parameter, and $\mathbf{R}_t$ stands for an estimate of regressors matrix of second moments, $E\left[\mathbf{x}_t \mathbf{x}_t'\right]$.

Algorithm 2 (SG)**.** *Under the estimation context of* (2.1)*, the* SG *algorithm assumes the form of*[5]

$$\hat{\boldsymbol{\theta}}_t^{SG} = \hat{\boldsymbol{\theta}}_{t-1}^{SG} + \mu_t \mathbf{x}_t \left(y_t - \mathbf{x}_t' \hat{\boldsymbol{\theta}}_{t-1}^{SG} \right), \tag{2.4}$$

where μ_t stands for the learning gain parameter.

Since the seminal works in the subject of learning and expectations in macroeconomics (e.g., Bray, 1982; Marcet and Sargent, 1989) the LS algorithm has been taken as the natural choice to represent agents mechanism of adaptive learning. This choice is in general attributed to the widespread knowledge of the LS as an estimator between econometricians. The SG algorithm, on the other hand, provides a computationally simpler alternative estimator, leading some authors to advocate for its use as a more plausible learning device from a bounded rationality standpoint (Barucci and Landi, 1997; Evans and Honkapohja, 1998b).

It is tempting to think of the SG as encompassed by the LS, simply setting $\mathbf{R}_t = \mathbf{I}$ and $\gamma_t = \mu_t$ in (2.2), and from such an observation to conclude that the SG, as a specific case of the LS, may

[4] This form is closer to that used in the adaptive learning literature under the name of Recursive Least Squares, for the case where the gain is decreasing with time, or Constant-Gain (Recursive) Least Squares, for the case of a time-invariant gain. In the engineering literature other variations in the nomenclature can be found and a computationally less demanding form is more common where the inversion of $\mathbf{R}_t$ in (2.2) is avoided by the use of the matrix inversion lemma (see Haykin, 2001).

[5] This form is common to both the adaptive learning (see also Evans et al., 2010) and the engineering literature. In the latter this algorithm is generally known as least mean squares, although commonly referring to the constant gain case, while stochastic gradient is often referred to the case of a time-decreasing gain. In some contexts, however, stochastic gradient is referred as a whole family of filters (see Macchi, 1995, p. 52).

not be expected to outperform this latter. It is one of our purposes here to qualify this assertion, by comparing the forecasting performance of these algorithms when applied to actual macroeconomic data. It is important to note, though, that from an *a priori* perspective (i.e., in theory), the above observation does not preclude the possibility that the simpler SG algorithm can outperform the encompassing LS specification[6].

In order to unify these two algorithms under a state-space representation we need to specify the law of motion for the time-varying coefficients of the linear model in (2.1). For that purpose we follow the recursive identification literature (Ljung and Soderstrom, 1983; Ljung and Gunnarsson, 1990), assuming a random walk hypermodel, i.e.,

$$\boldsymbol{\theta}_{j,t} = \boldsymbol{\theta}_{j,t-1} + \boldsymbol{\omega}_{j,t}, \ \forall j = 1, \dots, N. \tag{2.5}$$

The disturbance terms $\varepsilon_{j,t}$ and $\boldsymbol{\omega}_{j,t}$ are assumed to be zero mean mutually independent Gaussian[7] random sequences with variances (and covariances) given by $\sigma^2_{j,t} = E\left[\varepsilon^2_{j,t}\right]$ and $\boldsymbol{\Omega}_{j,t} = E\left[\boldsymbol{\omega}_{j,t}\boldsymbol{\omega}'_{j,t}\right]$, respectively. To allow independent estimation of each equation in (2.1) we also assume that there is no cross-correlation between their residuals, i.e., $E\left[\varepsilon_{i,t}\varepsilon_{j,t}\right] = 0 \, \forall i \neq j$.

Under this framework, we use the recent results of Berardi and Galimberti (2013) which established exact correspondences between the Kalman filter associated to the state-space estimation problem and the learning algorithms of our interest[8]. Also notice that these results permit straightforward extension of well established features associated to the Kalman filter to the context of the learning algorithms. One example is given by the use of the Kalman smoother for the initialization of these learning algorithms (see Berardi and Galimberti, 2012).

Finally, given our interest in unrestricted VAR specifications, we let $\mathbf{x}_t = (1, \mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-p})'$ where p denotes the VAR lag order and the unitary term refers to an intercept. Forecasts for the endogenous variables can be computed as usual,

$$\begin{aligned} \hat{\mathbf{y}}_{t+1|t} &= \hat{\boldsymbol{\theta}}'_t \mathbf{x}_{t+1|t}, \\ \hat{\mathbf{y}}_{t+2|t} &= \hat{\boldsymbol{\theta}}'_t \mathbf{x}_{t+2|t}, \\ &\dots \end{aligned} \tag{2.6}$$

where the subscripts on the variables indicate that forecasts are computed on the basis of previous period information, and $\hat{\boldsymbol{\theta}}_t = \left(\hat{\boldsymbol{\theta}}_{1,t}, \dots, \hat{\boldsymbol{\theta}}_{N,t}\right)$ concatenates the coefficients estimates. Also notice that $\mathbf{x}_{t+1|t} = (1, \mathbf{y}_t, \dots, \mathbf{y}_{t-p+1})'$, $\mathbf{x}_{t+2|t} = \left(1, \hat{\mathbf{y}}_{t+1|t}, \dots, \mathbf{y}_{t-p+2}\right)', \dots$, hence including only information available at period t.

[6] To corroborate this point, a brief literature review on *a prioristic* comparisons between the LS and the SG algorithms as estimators is presented in Appendix A.

[7] Gaussianity is not strictly necessary for our purposes, but is required to guarantee the optimality of the Kalman filter estimator associated to this non-stationary context.

[8] For convenience, we reproduce these correspondences in Appendix B.

2.3 Further implementation details

Projection facility

It is often of interest to restrict the coefficients for the law of motion in (2.1) to evolve within a bounded region in the parameter space. The requirement of stability for $\mathbf{y}_t$ is one of such cases, which under our VAR specification can be verified by using standard tools of multivariate time series analysis (see Lütkepohl, 2005, pp. 13-18). In the adaptive learning literature, this issue has been dealt with by resorting to the use of a device known as projection facility (Evans and Honkapohja, 1998a), which is a mechanism assumed to be coupled to the learning algorithms such that whenever the estimates leave a bounded region in the parameters space the device is activated in order to contain the escape.

There are different ways a projection facility can be conceived for implementation (see, e.g., Ljung and Soderstrom, 1983, pp. 366-368). One of the simplest, and the one we adopt, is to set the new coefficient estimates equal to their previous value whenever the new estimates obtained from the algorithm lead to a violation of the stability region. In spite of our use of a projection facility, this mechanism is only intended to control occasional data-led escapes of the estimates, and it should not be regarded as a main player in the algorithms' performance.

Upper bounds on learning gains

Beyond the properties of the data, it is the calibration of the learning gain parameters in the adaptive algorithms that mainly determines whether estimates will become unstable or not. Given the relevance of these parameters for the convergence and tracking performance of the algorithms, the determination of upper bounds[9] to form their range of admissible values has been a topic of great interest in the literature (see Haykin, 2001, and references therein).

Stability analysis reveals that the upper bound on the SG learning gain is dependent on the data, a result clearly connected to the finding that the SG algorithm is not scale invariant (see Evans et al., 2010). This is not the case for the LS algorithm though, and for this reason we treat the LS case first. Loosely speaking, the LS learning gain is usually allowed to assume any value between zero and one, i.e., $0 \leq \gamma_t < 1$. Notwithstanding, when implementing the LS algorithm with data, it is often the case that the unitary upper bound is not tight enough to ensure stability. Therefore, the upper bound for the LS gain, denoted by $\bar{\gamma}$, needs to be experimentally adjusted for the application of interest. This is done by departing from the unity value as a necessary, though not sufficient, initial upper bound.

The gain calibration gets much more involved for the case of the SG algorithm, due to the dependency of this algorithm's stability on the data. Haykin (2001) presents the derivation of a tight upper bound on the SG gain relying on an analysis of higher order modes of convergence. Without getting too much into

[9] The lower bound is given by zero for both LS and SG algorithms, where there will be no adaptation. Negative gain values are not reasonable as it would imply updating the coefficient estimates onto a direction opposite to that indicated by the observed error.

the details of this approach, this upper bound is given by $\bar{\mu}_K = {}^2/_{KS_{\max}}$, where K stands for the number of coefficients being estimated, and $S_{\max}$ stands for the maximum value of the power spectral density of the regressors[10]. Taking this upper bound as a reference, we proceed in the same way as we did for the LS calibration by adjusting it experimentally. The goal is to compute the SG algorithm over a range of gain values as wide as possible, but yet not incurring in a too frequent activation of the projection facility.

3 Design of Comparative Exercises

In this section we set forth the methodological approach we use to compare the plausibility of the LS and SG learning algorithms. We first propose and define two plausibility criteria that we will use in our study and we then discuss the model specification and the algorithms' calibration used for the computation of forecasts. Finally, we detail the data we use and describe a three-stage routine we adopt to empirically evaluate our plausibility criteria.

3.1 Plausibility criteria

In order to achieve our purpose of assessing the plausibility of different adaptive learning algorithms, we propose two comparative exercises, each designed departing from a distinct perspective on the use of the algorithms.

From the point of view of an economic agent, who has to build forecasts of variables relevant for economic decisions, what matters is the accuracy of such forecasts, as defined by their distance with actual future realizations. The plausibility criterion associated to this first exercise is therefore represented by the algorithms' *forecasting performance.*

From the point of view of the researcher, who is interested in uncovering which mechanism better represents the behavior of the economic agents being modeled, what matters is instead the resemblance of the forecasts produced by the algorithms to the observed forecasts as revealed from surveys. Thus, we succinctly denote the plausibility criterion associated to this exercise as *resemblance to surveys.*

Common to both exercises is the focus on evaluating the forecasts associated to the different learning algorithms, which is done using a loss function that measures their distance from a desired target.

Definition 1 (Loss function)**.** *The loss function $\mathcal{L}(\mathbf{f}(\mathbf{x}, \boldsymbol{\theta}), \mathbf{z})$ maps a set of outcomes, $\mathbf{z}$, and a set of forecasts, $\mathbf{f}$, to the real number line, where forecasts may depend on a set of parameters, $\boldsymbol{\theta}$, and observed variables, $\mathbf{x}$, and the loss function may accept multiple inputs, such as forecasts at different horizons*[11]*.*

[10] For applied purposes, this statistic requires estimation based on the regressors data. A summary of alternative methods for power spectrum estimation is presented in Haykin (2001, pp. 78-82). Here we use a simple periodogram applied to the whole sample of data.

[11] We are abstracting here from a possible dependency of losses on state variables, e.g., bigger losses during downturns of the business cycles (see Elliott and Timmermann, 2008).

The characterization of a loss function is typically made within the context of a decision problem. A relative ordering of different forecasts is obtained by comparing their expected losses, $E\left[\mathcal{L}|\mathcal{I}\right]$, where $\mathcal{I}$ stands for the conditioning information set under which the forecasts were computed and the losses evaluated. In our exercises, such conditioning involves two main aspects: the specification of a model and the choice of a gain calibration for the learning algorithms. Details about our assumptions on these points will be given below[12].

The main difference between our two exercises then rests on the outcomes, $\mathbf{z}$, being considered for the evaluation of the algorithms. Letting $\mathbf{y}$ stand for the set of values being forecasted, $\mathbf{f}_i$ for the set of forecasts associated to algorithm i, and $\mathbf{s}$ for the set of forecasts obtained from surveys, we define our two plausibility criteria as follows.

Criterion 1 (Forecasting Performance)**.** *An algorithm A is said to be a more plausible representation of agents'* OPTIMAL *behaviour in terms of learning-to-forecast than an alternative algorithm B if $E\left[\mathcal{L}\left(\mathbf{f}_A,\mathbf{y}\right)\right] < E\left[\mathcal{L}\left(\mathbf{f}_B,\mathbf{y}\right)\right]$.*

Criterion 2 (Resemblance to Surveys)**.** *An algorithm A is said to be a more plausible representation of agents'* ACTUAL *behaviour in terms of learning-to-forecast than an alternative algorithm B if $E\left[\mathcal{L}\left(\mathbf{f}_A,\mathbf{s}\right)\right] < E\left[\mathcal{L}\left(\mathbf{f}_B,\mathbf{s}\right)\right]$.*

3.2 Model and learning gain specifications

We will keep two things fixed across our exercises: the model specification agents are assumed to use for estimation and forecasting; and the way agents choose in real-time a gain calibration for the learning algorithms.

With respect to the model, our focus is on unrestricted VAR specifications applied to inflation and growth. We denote the time series for these variables by π_t and g_t, respectively. Under the notation of the previous section we then have $\mathbf{y}_t = \left(\pi_t, g_t\right)'$. For robustness we estimate VARs with lag orders from 1 to 4[13]. We shall discuss how our results are affected by the choice of this lag order, but with exception of some summarizing statistics, our presentation will focus on the results obtained with the VAR(1).

Regarding the gain calibrations, we need first to establish which options we consider available to agents, and then discuss how we assume agents choose among the available alternatives. On the first point, we follow the recent empirical literature on adaptive learning (see, e.g., Eusepi and Preston, 2011; Milani, 2011), and focus on specifications where the gains are kept constant throughout the sample, i.e., $\gamma_t = \gamma$ and $\mu_t = \mu$. Following our discussion in Section 2.3, we construct a grid of gain values by setting upper bounds on their admissible values so as to ensure the algorithm's stability. We use a grid of 100

[12] For our comparative purposes, we abstract from the use of $\mathcal{I}$ in the notation that follows.

[13] We do not allow these lag orders to change in real-time.

values for this purpose, meaning that the estimation routine that follows is applied to each algorithm with 100 different gain values[14].

When it comes to compare algorithms as learning-to-forecast devices, though, a choice of gain has to be made. In particular, real-time agents are required to specify a unique gain value each time an iteration on these recursive mechanisms is performed. Apart from being thought of as an agent's primitive learning parameter, the gain is also known to be prominent in determining the algorithm's tracking performance[15]. On the grounds of rationality, one may then argue that agents would be willing to optimise on their choice of a gain calibration, and this choice can therefore be viewed as determined by a minimization of the losses agents[16] expect to incur by using that gain to compute forecasts.

For our applied purposes, these expected losses are computed using sample averages. Hence, the choice of a gain calibration that minimises agents' expected losses turns into a problem of specifying a sample of forecasts from which average losses are computed. We explore three alternatives on this aspect:

Ex post: Pick the gain yielding the minimum average loss over the whole sample of forecasts that we have computed. This choice of gain clearly violates the restrictions of a "fair" out-of-sample forecasting exercise that we are exploring in connection to the idea of real-time learning. This alternative has also been used in some of the previous calibration attempts in the literature (see, e.g., Orphanides and Williams, 2005; Milani, 2007, 2008, 2011)[17].

Recursive: Allow the choice of the gain to be recursive, through a minimization of the average loss over a rolling window sample of forecasts. Here we adopt a window length of 60 periods of forecasts. From the real-time learning and forecasting perspective, this choice of gains is the extreme opposite to that proposed in the first alternative, as it does not allow the use of *ex post* information on the quality of the forecasts for the calibration of the algorithms.

In-sample: Split the whole sample of forecasts in two parts: the first part is used as an in-sample period on which the minimization of the average loss is applied in order to pick a gain for each algorithm; the second part is then used for the evaluation, keeping fixed the gain calibration. Here we also adopt an initial sample of 60 periods. This alternative goes in line with traditional exercises of forecasting evaluation and has also been used by Branch and Evans (2006) and Weber (2010) for purposes similar to ours.

[14] We have also computed the algorithms with a decreasing gain on the form of $\gamma_t = {}^{\bar{\gamma}}\!/_t$ and $\mu_t = {}^{\bar{\mu}}\!/_t$ to benchmark our results.

[15] See Appendix A.

[16] Clearly, we are talking about a loss in the form of that used in our *forecasting performance* plausibility criterion.

[17] We note that in all these papers the calibration has been made within a structural model and solely focusing on the LS. Further, in Orphanides and Williams (2005) the objective of the forecasts loss minimization was their resemblance to SPF's forecasts, not their performance.

3.3 Data

In order to define the setting in which forecasts can be interpreted as proxy for agents' macroeconomic expectations, it is important to be specific about the informational assumptions under which expectations are supposed to be formed. To deal with this issue, we adopt the environment provided by a real-time dataset (see Croushore, 2011)[18].

We use quarterly data on the US real GNP/GDP and its price index from 1947q2 to 2011q4, which sums up to 259 observations for each variable. Our data on these series come from the Philadelphia's Fed Real-Time Data Research Center[19] and consists of vintages from 1966q1 to 2012q1, i.e., a total of 185 snapshots of what was known on these variables by a market participant in real-time. As our interest is on forecasts for output growth and inflation, we obtain these rates from the above data on levels computing their associated annual growth rates by compounding their simple quarterly growth factors.

For the purpose of comparing the algorithms' forecasts to those provided by survey respondents, we use data from the Survey of Professional Forecasters (SPF)[20], which are made available by the Philadelphia's Fed as well. Each quarter, this survey asks professional economists to give their forecasts for several macroeconomic variables, including those we indicated above, and also over different forecasting horizons. Here we use the median of the individual forecasts for output growth and inflation made for horizons consisting of a total of five quarters ahead, namely from t to $t+4$.

3.4 Three-stage routine

Our approach to empirically evaluate the plausibility criteria defined above develops into three stages: initialization, estimation and forecasting, and evaluation. The first two are carried out identically for both of our two exercises. The different perspectives assumed for each of our plausibility criteria will unfold into different evaluative statistics in the last stage.

Initialization stage

The first step in the process of obtaining forecasts from the LS and SG algorithms is to set their initializations, which require estimates for $\hat{\boldsymbol{\theta}}_{j,0}$. Here we follow the smoothing approach of Berardi and Galimberti (2012), which requires a training sample of data to be left aside . For this purpose, we use the first 75 observations of our sample, namely, the data from 1947q2 to 1965q4. Devoting this amount of observations to the initialization stage goes in line with most of the empirical works that undertake

[18] For robustness, we have also analyzed forecasts obtained using revised (or historical) series of data, rather than the real-time vintages, and our main qualitative conclusions remained unaffected.

[19] See http://www.philadelphiafed.org/research-and-data/real-time-center/. We have done some specific adjustments to the original dataset, as detailed in Appendix C.

[20] See http://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/.

calibrations of adaptive learning rules of the kind we are dealing with here (see Berardi and Galimberti, 2012, and references therein).

Estimation and forecasting stage

With the initials given, we proceed to apply the LS and the SG algorithms to estimate VAR model specifications with the series of data on inflation and growth. Estimation and forecasting are carried out by vintage as follows:

1. The recursions for each algorithm/gain are computed departing from the vintage/algorithm/gain initials until exhaustion of the vintage sample.

2. The $t, \ldots, t+4$ forecasts for each vintage/algorithm/gain are computed using the last estimates of the model specification, where t stands for the vintage quarter.

We repeat these computations for each vintage of data from 1966q1 to 2010q4, which results in a total of 180 forecasts for each algorithm/gain/horizon both for inflation and growth.

Evaluation stage

At this stage we collect all the forecasts we have computed according to their different gain specification and compute their associated average losses according to the two different plausibility criteria defined above. Specifically, for our first exercise concerning the algorithms' *forecasting performance,* we compute average losses on the basis of forecast errors that measure the distance between each algorithm's forecasts and the actual realizations of the forecasted variables[21]. For our second exercise, comparing the algorithms' *resemblance to surveys*, average losses are based on the difference between each algorithm's forecasts and the forecasts obtained from surveys.

In both cases, a parametric form for the loss function still needs to be specified. Here we adopt three different forms: (i) the squared error, $\mathcal{L}_2(\mathbf{f}, \mathbf{z}) = (\mathbf{z} - \mathbf{f})' (\mathbf{z} - \mathbf{f})$; (ii) the absolute error, $\mathcal{L}_1(\mathbf{f}, \mathbf{z}) = \|\mathbf{z} - \mathbf{f}\|_1$; and, (iii) the fourth power of the error, $\mathcal{L}_4(\mathbf{f}, \mathbf{z}) = ((\mathbf{z} - \mathbf{f}) \circ (\mathbf{z} - \mathbf{f}))' ((\mathbf{z} - \mathbf{f}) \circ (\mathbf{z} - \mathbf{f}))$[22]. Our previous argument was that the specific characterization of the loss function is supposed to represent the preferences of the user of the forecasts. This would be the agent, in the first exercise, and the researcher, in the second. However, given our non-structural approach, it would be unconvincing at this stage to attempt to link these specifications to any well-founded preference framework. Instead, our use of these three different forms of loss function should be taken as an attempt to evaluate the robustness of our

[21] As for the vintage of data used as reference to compute these errors, for robustness we use both first-available and latest-available actuals (see Stark and Croushore, 2002). Given that our qualitative conclusions were found not to be affected by this choice, we present the results only for the former measure.

[22] Here $\|\bullet\|_1$ stands for the L_1 norm, i.e., the sum of the absolute values of the vector's elements, and $\circ$ stands for the Hadamard (entrywise) product.

results to varying conditions on the data environment[23].

To further substantiate our comparative analysis, we also make use of tests common to the literature on forecast evaluation. Namely, we adopt both the Diebold and Mariano (1995) (DM) test for equal (unconditional) predictive accuracy, and the more recently developed test for equal (conditional) predictive ability of Giacomini and White (2006) (GW). Other than for robustness purposes, our choice for these two tests can also be well justified: while the first stands as a classical test, whose properties have been long studied in the literature, the second clearly represents a more appropriate test for our purposes of comparison of different estimation methods. As its own proponents qualify, under the GW test "..the finite sample properties of the estimators on which the forecasts may depend are preserved asymptotically" (Giacomini and White, 2006, p. 1545), a feature of essential relevance for our empirical application.

4 Results

In this section we begin with an overview of the forecasting performance of each algorithm and then proceed to present results obtained under each of the two exercises comparing the algorithms' *forecasting performance* and *resemblance to surveys*. We end the section with a discussion of our findings.

4.1 Overview

One first result we can obtain about the LS and the SG algorithms' forecasts relates to the evolution of their performance over time for the sample at scrutiny, under different gain calibrations. This is presented in Figure 1 in the form of surfaces of average past performance for each algorithm and variable. Two main observations arise[24]: (i) the behavior of each algorithm depends on the variable being forecasted, whereas for a given variable the LS and SG algorithms behave differently; (ii) the magnitudes of forecast errors were relatively higher during the first decade in our sample, irrespective of the variable forecasted and the algorithm used. Rather than an indication of algorithms' instability, this latter observation can be associated with the period of greater volatility that preceded the Great Moderation in the US economy (see Stock and Watson, 2003).

In Figure 2 we aggregate these results through the time dimension, averaging forecast errors over our entire sample. Clearly, we can corroborate the observation of Branch and Evans (2006) that the LS with constant gain tends to outperform its decreasing gain version; furthermore, we also extend this result to the case of the SG algorithm. It is also evident in this figure how the performance of each algorithm is

[23] Clearly, taking the squared loss function as a benchmark, the absolute value loss function would tend to favor the algorithm performing better at periods of stability, where forecasting errors tend to be smaller than during periods of instability. The opposite can be said about the fourth power loss function.

[24] Inspection of such surfaces for different forecasting horizons, for errors computed using latest available data, and for the resemblance to surveys, bring up similar observations.

affected as the forecasting horizon varies. For the SG algorithm there is only a scale effect, where the magnitude of errors increases with the horizon. In contrast, for the LS algorithm the sensitivity of its forecasting performance to the gain calibration tends to increase at longer horizons.

This result is particularly relevant for the difficulty of finding support for the "learning explanation" of macroeconomic fluctuations among studies relying on RBC (Real-Business-Cycle) type of models. As the analysis of Eusepi and Preston (2011) reveals, learning has been introduced in these models directly into the Euler equations predicted by equilibrium and rational optimization assumptions. Elusively, that ends up turning the agent's infinite horizon decision problem into a one-period-ahead expectations formation problem. Our findings indicating the sensitivity of the algorithms' performance to variations in the forecasting horizon thus contribute to the relevance of Eusepi and Preston's argument.

4.2 Comparative exercises

Forecasting performance

We now evaluate our first plausibility criterion, comparing the *forecasting performance* of the LS and the SG algorithms within our pre-specified sample of data and framework design. Our analysis will be made numerically by pairing the average losses attained by each of these algorithms under the three different choices of gain calibration we discussed in Section 3.2: the *ex post*, the *recursive*, and the *in-sample* one. The mean squared forecast errors associated to each of these cases, together with their associated DM and GW pairwise testing statistics, are presented in Table 1. The idea here is to uncover not only which of these algorithms provides a better performance in forecasting these variables, but also to ascertain whether these results are statistically significant[25].

Our reading of the results in Table 1 is that they represent mixed evidence in support of each of the algorithms. Although the LS is in general outperforming the SG in the *ex post* choice of gain, none of these results are found with statistical significance. A different picture emerges from the *recursive* and *in-sample* choices of gain: the SG is found to outperform the LS in forecasting inflation most of the times, and the opposite is found for growth; some of these comparisons are found with statistical significance. It is in these latter two choices of gain that we think the agents' view is better represented. Thus, our conclusion is that the SG algorithm would have been the most plausible choice for agents' purposes of forecasting inflation, while the LS would take this place for the case of output growth forecasts.

The conclusion we have just drawn may well be put into question for its sensitivity with respect to the many specific aspects underlying the construction of the forecasts, such as the choice of the VAR lag order, or the loss function taken as representative of agents' preferences. In that respect, we pursued a sensitivity analysis, the results of which are synthesised in Table 2. The measure we use for that purpose is denoted as hit-rates, and refer to the frequencies by which each algorithm outperforms

[25] We will be using the 20% level of significance as our reference.

the other[26]. These were summed up over the 5 forecasting horizons and the 4 VAR lag orders we are evaluating. Clearly, we can see that the conclusions made above solely on the basis of the VAR(1) results are corroborated by the results we obtained under the higher lag orders.

We also evaluate how our conclusions are affected by the choice of the loss function. In Figure 3 we compare the hit-rates corresponding to the alternative specifications provided by the absolute value and the fourth power losses. Focusing on the results for the *recursive* and the *in-sample* choices of gain, we can see that the SG superiority we have observed in forecasting inflation is enhanced with the absolute value loss function. This indicates that the SG forecasts inflation better than the LS especially during usual times of stability. As for the superiority of the LS in forecasting growth, the opposite is observed, i.e., the SG performance in forecasting growth tends to improve relatively to that of the LS when a greater weight is given to larger errors.

Resemblance to surveys

To evaluate our second plausibility criterion we now compare the forecasts obtained from the LS and the SG algorithms to those we observe from surveys. Our analysis of the algorithms' *resemblance to surveys* follows an approach similar to that of our previous exercise. This allows us to go straight to the interpretation of the results, which are presented in Tables 3 and 4, and Figure 4.

In contrast to our previous findings, the results on *resemblance to surveys* indicate an overwhelming dominance of the LS algorithm. Other than beating the SG resemblance with a higher frequency, the victories of the LS are often found with statistical significance by the GW test. This observation is further corroborated by considering the different VAR lag orders. Also notice that, if the SG presents any threat to the LS dominance, this would not be on the grounds of their resemblance to inflation but to the growth survey forecasts. Clearly, this is a disturbing observation given that in our previous exercise we have found that the SG outperformed the LS in forecasting inflation, but not for growth.

Results obtained from varying the loss function show a similar pattern to the one we have observed in the previous exercise. Namely, for inflation the SG poses a threat to the LS dominance when a higher weight is given to smaller forecast comparison errors, i.e., using the absolute value loss function. The contrary is observed for the case of growth, where the success of the SG algorithm tends to be improved by the use of the fourth power loss function.

4.3 Discussion of results

An understanding of our results on the algorithms' forecasting performances can be obtained by considering the different statistical properties of the variables being forecasted. Compared to inflation rates,

[26] We focus on the GW statistics for its better suitability to our application. The DM test is well known to be undersized, especially for nested models (see Clark and West, 2006, between others), and this distortion seems to be manifest in our results.

output growth is known to have a lower degree of dynamic persistence and a higher degree of volatility. These characteristics naturally make output growth a variable harder to forecast than inflation (see, e.g., Patton and Timmermann, 2011). Hence, it seems reasonable to expect that a more sophisticated method would be favored in forecasting growth.

Our results corroborate this idea: the "sophisticated" LS tended to outperform the computationally simpler[27] SG algorithm in forecasting growth. Still, it is instructive to see that the SG was able to outperform the LS in forecasting inflation. These results thus confirm our preliminary view that the relative suitability of these algorithms depends mainly on the statistical properties of the data environment to which they are applied.

An intriguing result, nevertheless, emerged when we evaluated which of these algorithms provided forecasts closer to those we observe from surveys. Our evidence in that respect was found to favor the LS forecasts for both inflation and output growth variables. Obviously, taken alone such a result is not surprising, and can actually be taken to provide support for the usually unquestioned choice of the LS learning scheme in the applied literature on adaptive learning in macroeconomics[28]. The puzzling aspect of this result comes from its joint interpretation with those results we obtained for the performance comparison exercise.

The motivation we gave to these two exercises came from what one can think of as two distinguished pragmatic perspectives on the choice of a learning algorithm. For an economic agent, the requisite is of a mechanism capable of providing accurate inferences and, assuming this agent behaves rationally, such a choice would be guided by an optimization of forecasting performance. For a researcher, in contrast, the choice of a learning algorithm reflects his quest for the true mechanism representative of the former agent's behavior.

Our evaluation of the plausibility criteria associated to these two views therefore indicates a puzzling conflict in the choice of an algorithm representative of rational agents' learning. Rephrasing our findings in these interpretative terms, in order to form expectations for output growth and inflation: (i) neither the LS nor the SG algorithms dominate as what would have been the agents' rational choice; but, (ii) the LS was found to be the closest representative of agents' learning mechanism for both variables.

Finally, our analysis of the sensitivity of these results to different specifications of loss functions gives robustness to the existence of this conflict. Common sense dictates that larger errors can be attributed to rare events (or the tail of a distribution). If interest is on usual times (or the central modes of a distribution), then the focus should shift to a loss function attaching lower weights to larger errors (i.e., the absolute value loss). In these terms, our results indicate that the LS did a better job in tracking the central modes of the distribution of growth rates, whereas for the case of inflation rates the SG algorithm

[27] See Appendix A.

[28] As far as we are aware, the only study using the SG in a macroeconomic applied context is given by Bullard and Eusepi (2005).

assumed that position. We therefore conclude that our conflicting findings are not an artifact of rare (destabilizing) events, but a regularity of usual times of macroeconomic activity.

5 Concluding Remarks

Adaptive learning algorithms have been used to represent agents' process of expectations formation in macroeconomics, and the usually unquestioned choice for that purpose has been the LS algorithm. In this paper we have provided an empirical assessment on the plausibility of this choice by comparing the LS to its simpler competing alternative, the SG algorithm. Our comparative assessment was based on two distinct measures of plausibility, namely, their forecasting performance and their resemblance to survey forecasts.

We have motivated the use of these two plausibility criteria in terms of the pragmatic views of an agent and of a researcher. While the former is thought of as the ultimate user of the learning algorithms, the latter seeks to uncover which of the above algorithms provides the closer representation for agents' behavior of learning-to-forecast. Assuming the economic agents are rationally optimizing in their choice for a learning mechanism, we argue that from their perspective the most plausible algorithm is given by the one with a superior forecasting performance. From a researcher perspective, however, the most plausible option is the one that more closely resembles what is observed as agents' behavior, here represented by survey forecasts.

Combining these two perspectives, we obtained a conflicting answer to our main question on what is a plausible choice for a representative learning algorithm. While the rational choice of a learning algorithm was found to be dependent on the variable of interest, the LS algorithm was overwhelmingly favored as the algorithm providing forecasts closer to those of survey respondents. Although favorable to the main literature applying adaptive learning in the context of macroeconomic modeling, which often assumes the LS algorithm, this latter result conflicts with the former.

The puzzling aspect of these findings relates to the conflict they pose between what would be expected to be a rational choice of an optimizing agent and what we can infer from observed agents' behavior. It might seem tempting to take such conflict as evidence of a violation on the assumption of rational behavior. At the current stage, however, we warn against such explanation, and we hope instead that our findings will motivate future research in the direction of disentangling this conflict.

References

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

Benveniste, A., Metivier, M., Priouret, P., 1990. Adaptive Algorithms and Stochastic Approximations. Springer-Verlag.

Berardi, M., Galimberti, J.K., 2012. On the initialization of adaptive learning algorithms: A review of methods and a new smoothing-based routine. Discussion Paper Series 175. Centre for Growth and Business Cycle Research.

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

Branch, W.A., Evans, G.W., 2006. A simple recursive forecasting model. Economics Letters 91, 158–166.

Bray, M., 1982. Learning, estimation, and the stability of rational expectations. Journal of Economic Theory 26, 318–339.

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

Bullard, J., Eusepi, S., 2009. When does determinacy imply expectational stability? Working Papers 2008-007. Federal Reserve Bank of St. Louis.

Clark, T.E., West, K.D., 2006. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. Journal of Econometrics 135, 155–186.

Croushore, D., 2011. Frontiers of real-time data analysis. Journal of Economic Literature 49, 72–100.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253–263.

Elliott, G., Timmermann, A., 2008. Economic forecasting. Journal of Economic Literature 46, pp. 3–56.

Ellison, M., Pearlman, J., 2011. Saddlepath learning. Journal of Economic Theory 146, 1500–1519.

Eusepi, S., Preston, B., 2011. Expectations, learning, and business cycle fluctuations. American Economic Review 101, 2844–2872.

Evans, G., 1985. Expectational stability and the multiple equilibria problem in linear rational expectations models. The Quarterly Journal of Economics 100, 1217–1233.

Evans, G.W., Honkapohja, S., 1998a. Economic dynamics with learning: New stability results. Review of Economic Studies 65, 23–44.

Evans, G.W., Honkapohja, S., 1998b. Stochastic gradient learning in the cobweb model. Economics Letters 61, 333–337.

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

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

Eweda, E., 1994. Comparison of rls, lms, and sign algorithms for tracking randomly time-varying channels. Signal Processing, IEEE Transactions on 42, 2937–2944.

Eweda, E., 1999. Transient performance degradation of the lms, rls, sign, signed regressor, and sign-sign algorithms with data correlation. Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on 46, 1055–1062.

Giacomini, R., White, H., 2006. Tests of conditional predictive ability. Econometrica 74, 1545–1578.

Giannitsarou, C., 2005. E-stability does not imply learnability. Macroeconomic Dynamics 9, 276–287.

Hassibi, B., Kailath, T., 2001. H infinity bounds for least-squares estimators. Automatic Control, IEEE Transactions on 46, 309–314.

Hassibi, B., Sayed, A., Kailath, T., 1996. H infinity optimality of the lms algorithm. Signal Processing, IEEE Transactions on 44, 267–280.

Haykin, S.S., 2001. Adaptive Filter Theory. Prentice Hall Information and System Sciences Series, Prentice Hall, New Jersey, USA. 4th edition.

Heinemann, M., 2000. Convergence of adaptive learning and expectational stability: The case of multiple rational-expectations equilibria. Macroeconomic Dynamics 4, 263–288.

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

Ljung, L., Gunnarsson, S., 1990. Adaptation and tracking in system identification - a survey. Automatica 26, 7–21.

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

Lütkepohl, H., 2005. New Introduction to Multiple Time Series Analysis. Springer.

Macchi, O., 1995. Adaptive Processing: the least mean squares approach with applications in transmission. John Wiley & Sons.

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

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

McCallum, B.T., 2007. E-stability vis-a-vis determinacy results for a broad class of linear rational expectations models. Journal of Economic Dynamics and Control 31, 1376–1391.

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

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

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

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

Patton, A.J., Timmermann, A., 2011. Predictability of output growth and inflation: A multi-horizon survey approach. Journal of Business and Economic Statistics 29, 397–410.

Sargent, T.J., 1999. The Conquest of American Inflation. Princeton University Press, Princeton, NJ.

Sayed, A.H., 2008. Adaptive Filters. John Wiley & Sons, Hoboken, NJ.

Stark, T., Croushore, D., 2002. Forecasting with a real-time data set for macroeconomists. Journal of Macroeconomics 24, 507–531.

Stock, J.H., Watson, M.W., 1996. Evidence on structural instability in macroeconomic time series relations. Journal of Business and Economic Statistics 14, 11–30.

Stock, J.H., Watson, M.W., 2003. Has the business cycle changed and why?, in: NBER Macroeconomics Annual 2002, Volume 17. National Bureau of Economic Research, Inc. NBER Chapters, pp. 159–230.

Weber, A., 2010. Heterogeneous expectations, learning and European inflation dynamics. Cambridge University Press. chapter 12. pp. 261–305.

A Review of *a priori* comparisons between the algorithms

Computational complexity: The SG algorithm requires a lower number of computations for a complete iteration of adaptation than the LS, given that this latter requires the inversion of the matrix of moments, an operation for which computational complexity grows exponentially with the number of regressors. To be specific, while an SG iteration requires only $2K+1$ multiplications and $2K$ additions, a LS iteration requires K^2+5K+1 multiplications, K^2+3K additions, and 1 division, with K standing for the number of regressors in $\mathbf{x}_t$ (see Sayed, 2008, pps. 166, 200-201).

Rate of convergence: There are many factors affecting the rate of convergence of the LS and the SG algorithms, such as the level of noise and the eigenvalues of the regressors covariance matrix (data-driven factors); the magnitude and the direction of the initial misalignment in the coefficients estimates that first ignited the transient phase (misalignment factors); and, the magnitude of the learning gain for each algorithm (calibration factors). Eweda (1999) provides an analysis of the effects that these factors have over the algorithms' convergence time by focusing on simplified cases, i.e., by imposing restrictive assumptions on the data environment. One important factor is the learning gain calibration, which is found to determine the speed with which the algorithm estimates are adjusted in face of large misalignments: the higher is the learning gain, the quicker is the adjustment. A second observation from Eweda's (1999) results is that there might be cases when the SG algorithm converges faster than the LS algorithm. Specifically, under the case of uncorrelated regressors the SG transient performance tends to improve with increased regressors variance. Such a result is, however, dependent on the assumption of uncorrelated regressors, which is very restrictive for macroeconomic contexts. Under the case of correlated regressors, simulation evidence tends to favor the LS due to its data orthogonalization feature (see also Haykin, 2001, pps. 285-291, 367-371, 454-457).

Tracking performance: The assessment of an algorithm's tracking performance focuses on its steady state behavior, i.e., after passing the transient phase. One measure adopted for that purpose is the Mean-Square Deviation (MSD) between the actual vector of coefficients, $\boldsymbol{\theta}_t$, and the algorithms estimates, $\hat{\boldsymbol{\theta}}_t$, as given by $\mathcal{D}_t = E\left[\left\|\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t\right\|^2\right]$, which is intended to capture the (average) accuracy of the algorithm's estimates. Its evolution through time is also associated with the speed with which the algorithm is able to adjust its estimates to the time-varying system, and optimization of tracking performance is in general associated to a minimization of MSD, mainly through control of the gain parameter. In attempting to do so, however, one is confronted with a well known trade-off between speed and accuracy in the estimates provided by the algorithms: on one extreme, tracking can be slower than the system actual time variations, but with less noisy estimates; on the other extreme, tracking can be made as rapid as the time-varying context, but with estimates much more contaminated by noise (see e.g. Benveniste et al., 1990, Part I, Chapters 1 and 4). Comparative analysis on the tracking performance of the LS and the SG algorithms can be found in Eweda (1994) and Haykin (2001, pp. 643-659), and their results indicate the preeminence of data conditions in the determination of which algorithm outperforms the other. To be more specific, the comparison between the LS and the SG algorithm in terms of tracking performance depends on how the covariance matrices of the regressors (say $\mathbf{x}_t$ in (2.1)) and of the disturbances affecting the time-varying coefficients (i.e., $\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}$) relate to each other.

Robustness: In a context of model misspecification, an estimator is said to be robust if it does not magnify the effect of modeling errors on estimation errors, and the SG algorithm is known to be the maximally robust algorithm in this sense (see Hassibi et al., 1996; Evans et al., 2010, pp. 240-242). Loosely speaking, in a worst case of misspecification the magnitude of the prediction errors obtained from the SG estimation will never exceed the magnitude of the true model disturbances. The LS algorithm, in turn, does not have this same robust interpretation, a result that can be associated to its squared norm formulation (Hassibi and Kailath, 2001). In short, the robustness of the LS algorithm, as measured by the (median) level of the interval of possible factor values at which modeling errors are propagated into estimation errors, depends on statistics of the regressors data ($\mathbf{x}_t$ in (2.1)).

B Correspondences between learning algorithms and Kalman filter

Adapted to the context of (2.1)-(2.5) the Kalman filtering recursion is given by

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{j,t} &= \hat{\boldsymbol{\theta}}_{j,t-1} + \mathbf{K}_{j,t}\left(y_{j,t} - \mathbf{x}_t'\hat{\boldsymbol{\theta}}_{j,t-1}\right), && \text{(B.1)}\\
\mathbf{K}_{j,t} &= \frac{\mathbf{P}_{j,t-1}\mathbf{x}_t}{\mathbf{x}_t'\mathbf{P}_{j,t-1}\mathbf{x}_t + \sigma_{j,t}^2}, && \text{(B.2)}\\
\mathbf{P}_{j,t} &= \left(\mathbf{I} - \frac{\mathbf{P}_{j,t-1}\mathbf{x}_t\mathbf{x}_t'}{\mathbf{x}_t'\mathbf{P}_{j,t-1}\mathbf{x}_t + \sigma_{j,t}^2}\right)\mathbf{P}_{j,t-1} + \boldsymbol{\Omega}_{j,t}, && \text{(B.3)}
\end{aligned}$$

where $\mathbf{K}_{j,t}$ is known as the Kalman gain vector and $\mathbf{P}_{j,t}$ stands for the covariance matrix of the coefficients estimates, i.e., $\mathbf{P}_{j,t} = E\left[\left(\boldsymbol{\theta}_{j,t} - \hat{\boldsymbol{\theta}}_{j,t}\right)\left(\boldsymbol{\theta}_{j,t} - \hat{\boldsymbol{\theta}}_{j,t}\right)'\right]$. The LS algorithm can be obtained as the special case of the Kalman filter by setting

$$\begin{aligned}
\sigma_{j,t}^2 &= \frac{\gamma_{j,t-1}}{\gamma_{j,t}}\left(1 - \gamma_{j,t}\right), && \text{(B.4)}\\
\boldsymbol{\Omega}_{j,t} &= \left(\frac{1 - \sigma_{j,t}^2}{\sigma_{j,t}^2}\right)\left(\mathbf{I} - \frac{\mathbf{P}_{j,t-1}\mathbf{x}_t\mathbf{x}_t'}{\mathbf{x}_t'\mathbf{P}_{j,t-1}\mathbf{x}_t + \sigma_{j,t}^2}\right)\mathbf{P}_{j,t-1}, && \text{(B.5)}
\end{aligned}$$

while the SG algorithms can be found as the special case of the Kalman filter when we set

$$\sigma^2_{j,t} = \mu^{-1}_{j,t} - \mathbf{x}'_t\mathbf{x}_t, \tag{B.6}$$

$$\mathbf{\Omega}_{j,t} = \mathbf{I} - \left(\mathbf{I} - \frac{\mathbf{P}_{j,t-1}\mathbf{x}_t\mathbf{x}'_t}{\mathbf{x}'_t\mathbf{P}_{j,t-1}\mathbf{x}_t + \sigma^2_{j,t}}\right)\mathbf{P}_{j,t-1}. \tag{B.7}$$

Details on these correspondences can be found in Berardi and Galimberti (2013).

C Details on data

Short time series history: some vintages lack of earlier observations due to delays into BEA revisions (see Philadelphia's Fed documentations). This was the case of the vintages of 1992q1-1992q4 (missing data from 1947-1958), 1996q1-1997q1 (missing data from 1947-1959q2), and 1999q4-2000q1 (missing data from 1947-1958). We circumvent this problem (to turn the dataset vintages-balanced) by reproducing observations from the last available vintage while rescaling in accordance to the ratio between the first observation available in the missing observation vintage and the value observed for the same period in the vintage being used as source for the missing observations.

Missing observation for 1995q4 in vintage 1996q1: as a result of the US federal government shutdown in late 1995, the observation for 1995q4 was missing in the 1996q1 vintage. Fortunately, this is the only point in this dataset that this happens. We fulfill this gap by using the observation available in the March 1996 monthly vintage for the same series. Incidentally, the SPF 1996q1 median backcast for 1995q4 is identical to the value later observed in March 1996, thence, our simplifying procedure is not favoring any method.

Caveat on SPF's forecasts for Real GDP: forecasts for real GDP were not asked in the surveys prior to 1981q3. To extend this series of forecast back to 1968q4, real GDP prior to 1981q3 is computed by using the formula (nominal GDP / GDP prices) * 100.

D Tables

Table 1: Forecasts *performance* comparative evaluation table under squared loss.

Gain choice	Forecasting horizons for INFLATION					Forecasting horizons for GROWTH				
- Statistics	0	1	2	3	4	0	1	2	3	4
Ex post (180 forecast obs.)										
- LS average loss	**2.10**	**3.02**	**3.42**	3.84	**4.60**	**7.72**	**9.90**	**10.83**	**10.69**	**10.55**
- SG average loss	2.36	3.38	3.56	**3.59**	4.75	8.77	11.18	11.94	12.24	12.34
- DM-statistic	-0.13	-0.09	-0.02	0.04	-0.02	-0.08	-0.09	-0.09	-0.12	-0.14
- DM-p-value	0.90	0.93	0.98	0.97	0.99	0.93	0.93	0.93	0.90	0.89
- GW-statistic	0.00	0.01	0.53	0.18	0.45	1.16	0.08	0.02	0.01	0.03
- GW-p-value	0.99	0.94	0.47	0.67	0.50	0.28	0.78	0.89	0.92	0.87
Recursive (120 forecast obs.)										
- LS average loss	1.69	1.88	1.97	2.32	2.74	5.55	6.99	**7.25**	**6.29**	**5.76**
- SG average loss	**1.69**	**1.86**	**1.26**	**1.34**	**1.95**	**5.38**	**6.94**	7.37	7.66	7.12
- DM-statistic	0.00	0.01	0.15	0.19	0.12	0.03	0.00	-0.01	-0.13	-0.12
- DM-p-value	1.00	0.99	0.88	0.85	0.91	0.98	1.00	0.99	0.89	0.90
- GW-statistic	0.90	0.24	1.70	0.02	0.16	2.92	0.37	2.33	2.34	0.09
- GW-p-value	0.34	0.62	**0.19**	0.89	0.69	**0.09**	0.54	**0.13**	**0.13**	0.76
In-sample (120 forecast obs.)										
- LS average loss	**1.64**	1.96	1.84	2.16	2.60	5.72	**6.69**	**6.16**	**6.16**	**5.57**
- SG average loss	1.76	**1.86**	**1.22**	**1.42**	**2.01**	**5.52**	7.30	7.82	8.18	7.91
- DM-statistic	-0.07	0.04	0.13	0.14	0.09	0.03	-0.06	-0.16	-0.18	-0.18
- DM-p-value	0.94	0.97	0.90	0.89	0.93	0.98	0.95	0.87	0.86	0.85
- GW-statistic	0.03	0.48	1.84	0.01	0.01	2.68	0.13	0.27	0.71	0.07
- GW-p-value	0.87	0.49	**0.18**	0.91	0.91	**0.10**	0.72	0.60	0.40	0.80

Forecast errors are computed with respect to first-available observation in the real-time dataset, while the forecasts itself are computed solely on the basis of real-time available observations with a VAR(1) specification. In the comparison of average loss within each choice of gain (see the text for explanations), we highlight in **bold** the algorithm presenting lower loss by horizon/variable forecasted. We do the same to highlight those comparisons for which statistically significant differences between the forecasts provided by each algorithm is found at levels below 20%, using the Diebold and Mariano (1995) (DM) and the Giacomini and White (2006) (GW) tests.

Table 2: Forecasts *performance* hit-rates summing up forecasting horizons (5) and VAR lag orders (4) under squared loss.

Variables	LS beats SG		SG beats LS	
- Gain choice	Hit rate	GW-20%	Hit rate	GW-20%
Inflation				
- Ex post	**70%**	20%	30%	10%
- Recursive	35%	0%	**65%**	20%
- In-sample	20%	0%	**80%**	20%
Growth				
- Ex post	**95%**	5%	5%	0%
- Recursive	**75%**	40%	25%	15%
- In-sample	**80%**	10%	20%	5%

Hit-rate here stands for the frequency within which one algorithm achieves a lower average loss than (= beats) the other algorithm in forecasting the indicated variables. These measures are computed on the basis of results of the kind presented in Table 1, with one such a table constructed for each VAR lag order. The count of beats favoring each of the algorithms is then averaged over the 5 forecasting horizons of $h = t, \ldots, t+4$, and 4 VAR lag orders $(1, \ldots, 4)$, and GW-20% represents the frequency within which the associated beats were found with statistical significance below the 20% level. See also the notes to Table 1, and the text for the explanations of each choice of gain.

Table 3: Forecasts *resemblance* comparative evaluation table under squared loss.

Gain choice	Forecasting horizons for INFLATION					Forecasting horizons for GROWTH				
- Statistics	0	1	2	3	4	0	1	2	3	4
Ex post (180 forecast obs.)										
- LS average loss	**0.86**	**0.93**	**1.00**	**0.86**	**1.06**	**3.83**	**3.54**	**2.39**	**1.28**	**1.04**
- SG average loss	0.96	1.02	1.15	1.31	1.56	3.87	4.47	4.41	4.40	4.46
- DM-statistic	-0.08	-0.03	-0.05	-0.10	-0.08	-0.01	-0.16	-0.26	-0.43	-0.46
- DM-p-value	0.93	0.97	0.96	0.92	0.94	1.00	0.87	0.80	0.67	0.65
- GW-statistic	2.49	0.01	2.24	0.10	2.16	4.93	0.69	14.33	22.48	17.29
- GW-p-value	**0.11**	0.94	**0.13**	0.76	**0.14**	**0.03**	0.41	**0.00**	**0.00**	**0.00**
Recursive (120 forecast obs.)										
- LS average loss	**0.56**	**0.47**	**0.33**	**0.29**	**0.42**	2.61	**1.62**	**1.14**	**0.64**	**0.54**
- SG average loss	0.63	0.78	0.89	1.06	1.36	**2.23**	2.08	2.25	2.59	2.25
- DM-statistic	-0.08	-0.16	-0.24	-0.21	-0.18	0.12	-0.07	-0.15	-0.27	-0.24
- DM-p-value	0.94	0.87	0.81	0.84	0.86	0.91	0.95	0.88	0.79	0.81
- GW-statistic	1.79	0.00	2.92	3.18	2.04	1.17	3.64	7.01	7.49	7.11
- GW-p-value	**0.18**	0.94	**0.09**	**0.07**	**0.15**	0.28	**0.06**	**0.01**	**0.01**	**0.01**
In-sample (120 forecast obs.)										
- LS average loss	**0.55**	**0.49**	**0.45**	**0.42**	**0.40**	2.85	**1.59**	**0.70**	**0.55**	**0.42**
- SG average loss	0.62	0.75	0.79	0.99	1.21	**2.20**	2.48	2.92	3.44	3.42
- DM-statistic	-0.09	-0.14	-0.17	-0.16	-0.15	0.14	-0.16	-0.39	-0.49	-0.55
- DM-p-value	0.93	0.89	0.87	0.87	0.88	0.89	0.87	0.70	0.62	0.59
- GW-statistic	2.01	0.01	2.30	0.75	1.85	2.04	8.04	14.12	16.44	20.23
- GW-p-value	**0.16**	0.91	**0.13**	0.39	**0.17**	**0.15**	**0.00**	**0.00**	**0.00**	**0.00**

Forecast comparison errors are computed with respect to the SPF's forecasts for the corresponding horizons, while the algorithm's forecasts itself are computed solely on the basis of real-time available observations with a VAR(1) specification. In the comparison of average loss within each choice of gain (see the text for explanations), we highlight in **bold** the algorithm presenting lower loss by horizon/variable forecasted. We do the same to highlight those comparisons for which statistically significant differences between the forecasts provided by each algorithm is found at levels below 20%, using the Diebold and Mariano (1995) (DM) and the Giacomini and White (2006) (GW) tests.

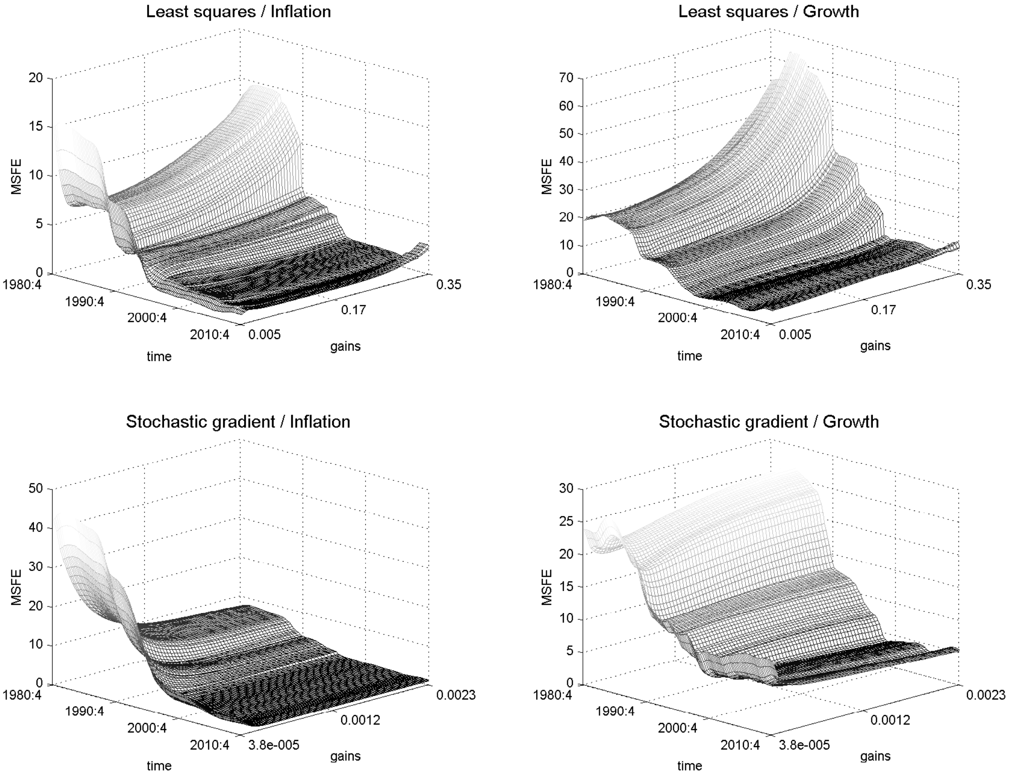
Table 4: Forecasts *resemblance* hit-rates summing up forecasting horizons (5) and VAR lag orders (4) under squared loss.

Variables	LS beats SG		SG beats LS	
- Gain choice	Hit rate	GW-20%	Hit rate	GW-20%
Inflation				
- Ex post	**80%**	20%	20%	0%
- Recursive	**100%**	40%	0%	0%
- In-sample	**80%**	30%	20%	5%
Growth				
- Ex post	**80%**	65%	20%	15%
- Recursive	**65%**	65%	35%	5%
- In-sample	**75%**	70%	25%	20%

See the notes to Table 2, only noting that the present table hit-rates are computed on the basis of results of the kind presented in Table 3 on the forecasts resemblance.

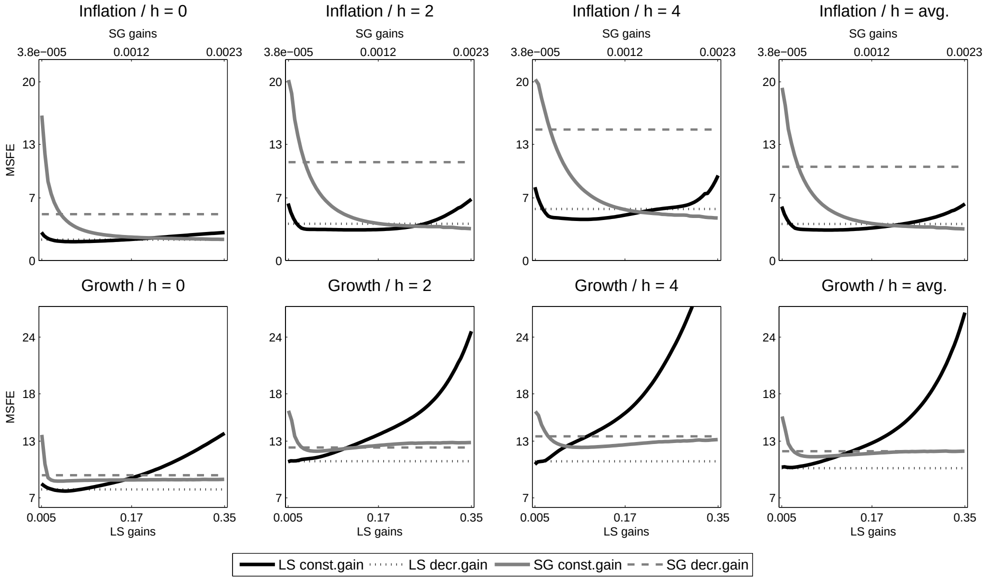
E Figures

Figure 1: Evolution of algorithms' forecasting performance through time and according to the learning gain, by variable forecasted.



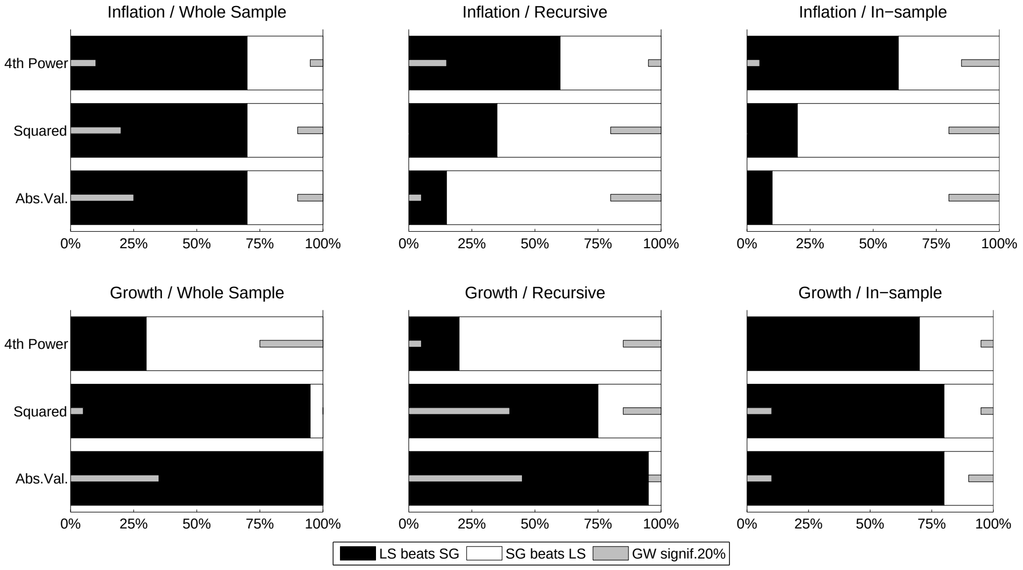
The mean squared forecast errors (MSFE) plotted are computed on the basis of the 60 quarters backwards to the dates indicated into the NW-SE axis. The gain calibrations used for each algorithm are indicated into the SW-NE axis, and correspond to those experimentally calibrated for the algorithms' stability. Forecast errors refer to mean-aggregated errors over 5 horizons of forecasts, and are computed with respect to first-available observation in the real-time dataset. The forecasts are computed solely on the basis of real-time available observations with a VAR(1) specification.

Figure 2: Algorithms' forecasting performance aggregated over time according to the learning gain, by variable forecasted and forecasting horizon.



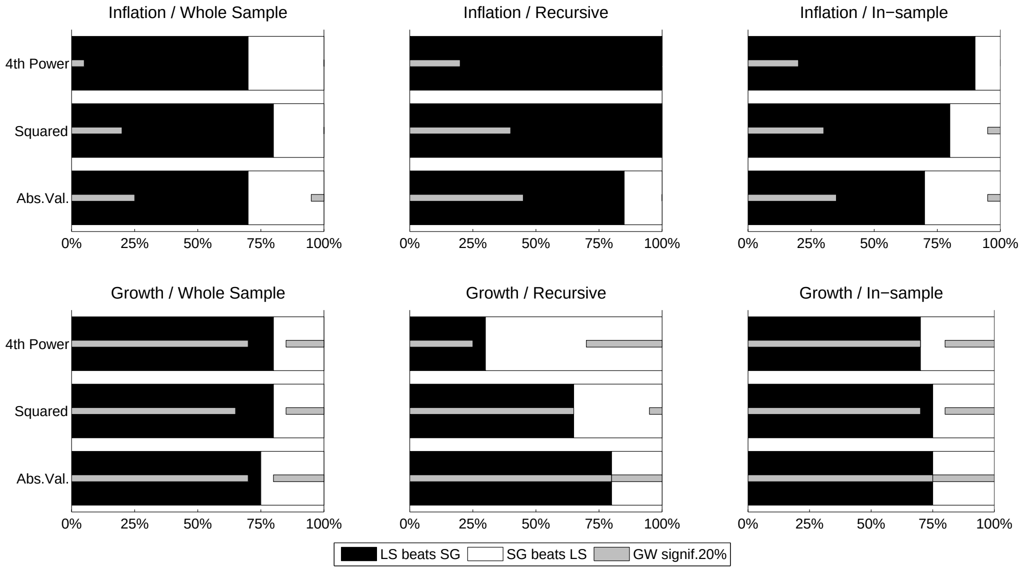
The mean squared forecast errors (MSFE) plotted are computed for each gain over the whole sample of forecasts, 1966q1-2010q4. The gain calibration used for the LS/SG algorithms are indicated into the lower/upper horizontal axes, respectively, and correspond to those experimentally calibrated for algorithm's stability. The last horizon plotted for each variable refers to that of mean-aggregated errors over 5 horizons of forecasts. All errors are computed with respect to first-available observation in the real-time dataset. The forecasts are computed solely on the basis of real-time available observations with a VAR(1) specification.

Figure 3: Forecasting *performance* hit-rates with respect to the loss function, by variable and choice of gain.



The hit-rates plotted follow the same definition as used in Table 2, but here for varying loss functions (see the text in Section 3.4). Note, for instance, that the squared loss function benchmark (middle bars) plots the same measures represented into Table 2.

Figure 4: *Resemblance* to surveys hit-rates with respect to the loss function, by variable and choice of gain.



The hit-rates plotted follow the same definition as used in Table 4, but here for varying loss functions (see the text in Section 3.4). Note, for instance, that the squared loss function benchmark (middle bars) plots the same measures represented into Table 4.