1 Introduction

I have been doing research on expectations in macroeconomics for twenty years. When I started, back in the year 2000, almost every model assumed rational expectations. There were no alternative assumptions that were simultaneously (i) tractable across models, (ii) consistent within each model, and (iii) with few parameters to set. At the same time, most empirical studies of survey data rejected the null hypothesis of rational expectations. In the data, people’s stated forecast errors turned out to be sometimes biased, often persistent, and always inefficient.

At the time, it felt that progress required new models to fill this gap. So, this is what I did, writing models of sticky information and inattentiveness that only had one parameter to calibrate, that could be inserted as assumptions in any model of dynamic decisions, and that were as easy to solve as models with rational-expectations (Mankiw and Reis, 2002, 2010). Many others were in the same pursuit, and in these two decades the theoretical literature has flourished with models of expectation that are as good or better in satisfying these criteria, including: dispersed private information (Woodford, 2003, Angeletos and Lian, 2016), rational inattention (Sims, 2003, Mackowiak, Matejka and Wiederholt,

Just as impressive has been the progress on empirically analyzing survey data on expectations. Research moved far beyond just computing measures of central tendency in the data, or just testing the null hypothesis of rational expectations. Exciting new results have come from looking at the dynamics of expectations (Coibion and Gorodnichenko, 2012, Andrade et al., 2019), their revisions and reaction to news (Coibion and Gorodnichenko, 2012, Bordalo et al., 2018), disagreement within surveys (Mankiw, Reis and Wolfers, 2004), disagreements across surveys (Carroll, 2003), information and regime treatments (Capistran and Ramos-Francia, 2010, Coibion, Gorodnichenko and Weber, 2019), differences across horizons (Andrade et al., 2016), uncertainty (Binder, 2017) and the link from expectations to actions (Bachmann, Berg and Sims, 2015, Coibion, Gorodnichenko and Ropele, 2020) and to inflation dynamics (Coibion, Gorodnichenko and Kamdar, 2018).

Today, in 2020, we have a wealth of insights from this literature. Researchers are making progress across many dimensions and, understandably, they spend their energy debating the often-subtle ways in which these models differ and marginally improve our understanding. From the perspective of outsiders, however, what stands out too often is a bewildering wilderness of alternatives. Today, few wince at a researcher making an alternative assumption on expectations in a seminar, but at the same time, few also teach anything but rational expectations in a core macroeconomics class. The core knowledge that gets passed in textbooks and classes consists of some key insights in a few parsimonious models. The literature on non-rational expectations has not yet produced its own basic model.

A comparison with two other building blocks of macroeconomic models may help clarify what I mean by this. Every macroeconomist knows that the Cobb-Douglas production function is wrong. It is easy to reject the null hypothesis that one can aggregate multiple inputs into only capital and labor, and the elasticity of substitution between them is surely not constant, let alone equal to one. Yet, decades of research on production and technology have convinced most that the Cobb-Douglas specification is a good starting
point. Wrong, yes, but useful, capturing fundamental principles of a technology frontier, of substitution between factors, or of the link between average and marginal products. When a researcher sits to write a paper focusing on expectations, she feels confident in assuming a Cobb-Douglas function in the production side of the economy, aware that she is missing some features, but confident that she is capturing the basics of production. The same could be said for assuming expected utility, a constant relative risk aversion utility function, or Calvo price rigidity. We have learned, through hundreds of research papers, that these baseline specifications capture basic features of behavior that are fundamentally important. The non-rational expectations literature is missing such an off-the-shelf model.

2 Yes, one parsimonious model of expectations

Angeletos, Huo and Sastry (2020) propose such a parsimonious model. Consider only macroeconomic models that are linear and stationary, in the sense that their endogenous variables $z_t$ depend on a set of exogenous disturbances $\varepsilon$ according to a Wold representation:

$$z_t = R(L)\varepsilon_t. \tag{1}$$

The lag polynomial $R(L) = R_0 + R_1 L + R_2 L^2 + \ldots$ is the solution of the model, with sparse $R_i$ matrices whose many elements depend only a few economic parameters. Most dynamics macroeconomic models fit into this description.

Equilibrium is, as always, a fixed point between beliefs and outcomes. For concreteness assume that there is a continuum of individuals $i$ that each form some subjective belief on some or all of the macroeconomic variables: $\hat{E}_{i,t}[z_t]$. At this level of generality, just write equilibrium as:

$$z_t = f\left(\{\hat{E}_{i,t}[z_t]\}_{i \in [0,1]}\right), \tag{2}$$

where $f(\cdot)$ is the function mapping agents’ expectation to their behavior and to market clearing conditions, which in turn determine the actual macroeconomic outcomes.

The non-rationality of expectations comes from two ingredients. First, agents in the model perceive macroeconomic dynamics to be given by a different model that has $\hat{R}(L) \neq R(L)$. Many behavioral biases can map into different specifications for this perceived process. For instance, agents that over-extrapolate into the future are those that perceive the variable to be more persistent than what it is. Agents that use heuristics may neglect
cross-correlations, but perceive each macroeconomic variable as being a univariate process. Agents that are scarred by their younger formative years may put a larger Wold weight on the disturbances realised during their twenties.

Second, conditional on this misperceived process, agents receive noisy individual signals of the macroeconomic variables and use Bayesian signal extraction to form their expectation. In particular, agents observe \( z_t + \tau^{-1/2}u_{i,t} \) where the \( u_{i,t} \) is a vector of idiosyncratic noises, each of of mean zero and unit variance, so that \( \tau \) measures the inverse precisions of these signals. This captures the incomplete and dispersed nature of information that comes out of the literatures on inattention. Angeletos, Huo and Sastry (2020) also explore the possibility that the actual precision \( \tau \) may be lower than the perceived precision, namely because agents may be overconfident. This does not seem to play a large role, according to their estimates, so I ignore it.

Letting \( E_t(\cdot) \) denote a Bayesian expectation, then the parsimonious model of beliefs is:

\[
\hat{E}_{i,t}[z_t] = E_t[z_t | \hat{R}(L) + \tau^{-1/2}u_{i,t}].
\]

This is a promising setup. It is simple, easy to explain, and flexible.

As stated above, this approach still has too many free parameters. But, with some discipline so that \( R(L) - \hat{R}(L) \) depends only on a couple of parameters, and a reasonable restriction that \( \tau \) is diagonal, then one gets closer to the desired. For instance, consider a very simple flexible-price model where inflation follows an AR(1) in equilibrium with parameter \( \rho \), and agents receive noisy signals on its realizations. Then, there are only two expectational parameters: the perceived persistence \( \hat{\rho} \) and the precision of the signals \( \tau \). Moreover, the rich empirical literature has provided strong guidance on how to set these two parameters.

### 3 Evidence on over and under reaction, and sluggishness

A simple panel-data regression nicely captures two of the main insights of the literature that has looked at survey expectations over the last decade. Take the case where \( z_t \) is a scalar, namely inflation, to take advantage of the available good long panels of data on
inflation expectations. Then, define the following variables:

\[
Error_{i,t} = z_{t+1} - \hat{E}_{i,t} [z_{t+1}]
\]  
\[
Revision_{i,t} = \hat{E}_{i,t} [z_{t+1}] - \hat{E}_{i,t-1} [z_{t+1}]
\]  
\[
AvRevision_t = \int Revision_{i,t} di
\]  

The regression is:

\[
Error_{i,t} = \kappa AvRevision_t - \chi (Revision_{i,t} - AvRevision_t) + u_{i,t}
\]  

With monthly or quarterly data on a panel of people reporting their expectations of inflation over the next year, as we for instance have in the Survey of Professional Forecasters, we can estimate this regression.

The typical estimates are \( \kappa > 0 \) and \( \chi < 0 \), and they reveal two salient features of the data. First, imagine one averaged both sides of the regression equation across people. The \( \chi < 0 \) would drop out, and the regression of average forecast errors on average forecast revisions would capture the stickiness of expectations. When a shock raises inflation, people, on average, increase their expectations by less than the new reality. The positive \( \kappa \) reflects this under-reaction over time. It produces a positive serial correlation of forecast errors, a fact that study after study has found in the data (Coibion and Gorodnichenko, 2015).

Second, imagine that one only had cross-sectional data to estimate this equation, so that only the \( \chi \) was identified. A negative estimate then indicates that those that revise their expectations by more, over-do it, and so end up making forecast errors in the opposite direction. A negative \( \chi \) captures the over-reaction in the cross-section, that may be attributable to over-confidence on current data being over-representative of what the future will be like. The data on forecasts of financial variables, in particular, shows strong evidence of this behavior (Bordalo et al., 2018).

The literature has often struggled with these two facts, with some studies finding over-reaction, and others under-reaction. The panel regression makes clear that these apparently disparate results are explained by either leaving out one of the two variables on the right-hand side, or by having the variation in the data be dominated by the cross-section or the time-series dimension. People on average under-react over time but, conditionally on that, individually overreact in the cross-section.
Macro models are often evaluated not by regression coefficients, but rather by their Wold representation $R(L)$. Figure 1 shows it for inflation, where, as in Angeletos, Huo and Sastry (2020), I use a particular reduced-form shock that accounts for a large share of business-cycle variation in a few macroeconomic series. The estimates of these impulse response function come from local projections and data on the GDP deflator between 1968:Q4 and 2017:Q4. Inflation follows familiar sluggish and hump-shaped dynamics after a shock.

The figure shows also the impulse response of survey forecasts using the average forecast in the Survey of Professional Forecasters for inflation on year ahead. It is important to be precise about what this measures. It is the change in the subjective forecasts of agents, as expected by the econometrician forming rational expectations according to her statistical model. These are not the actual changes in those subjective expectations. Mathematically:

$$E_t \left[ \frac{\partial}{\partial \varepsilon_t} \hat{E}_{t,t+h}(z_{t+h+4}) \right] \neq \frac{\partial}{\partial \varepsilon_t} E_t(z_{t+h+4})$$

(8)

because the law of iterated expectations need not hold across the expectations operators.

The sluggishness of expectations is clear. Following a shock, expectations are sticky, only catching up to reality 8 quarters after the shock. Forecast errors are positive not just
on impact, but for a prolonged period. It is this stickiness of expectations that, in varied ways, the models of the last two decades have tried to make sense of.

4 The delayed overshooting of expectations

Angeletos, Huo and Sastry (2020) highlight another feature of figure 1. The impulse response of forecasts crosses that of the actual variable from above after 8 quarters, and stays below it afterwards. While average forecasts errors are positive initially, they become negative after some periods. This reversal of the sign of the forecast error pins down a very particular pattern for the expectational parameters. The initial sluggishness reflects the noisy and imperfect information that people have on current shocks. The later negative forecast errors reflect an over-extrapolation that makes people expect the shocks’ effect to persist longer than they do. In terms of the simple AR(1) model:

\[ z_t = \rho z_{t-1} + \epsilon_t \]

where agents perceive \( \hat{\rho} \) and get signals \( z_t + \tau^{-1/2} u_i,t \), the initially positive forecast errors point to a small \( \tau \) while the later negative errors point to \( \hat{\rho} > \rho \).

This proposed fact is promising. However, for now, to my eyes, it is only suggestive, rather than definitive. Two points give me pause before taking figure 1 and its many variants that Angeletos, Huo and Sastry (2020) put forward as establishing delayed overshooting of expectations as a solid fact. First, the standard errors associated with these impulse response functions are wide, especially at longer horizons. Statistical tests at conventional significance levels can reject the null hypothesis that the forecast errors are zero in the first few quarters, against the alternative that they are positive. But, rejecting the null hypothesis of zero for longer horizons is much harder. From a Bayesian perspective, if one starts with a prior for the average forecast error at horizons 10 to 20 centered at zero, the evidence in figure 1 and in Angeletos, Huo and Sastry (2020) would tilt the posterior towards being more inclined to the forecast errors being negative rather than positive, but not by a lot. The difference between the two curves in figure 1 after 10 quarters is not so large, and the estimation uncertainty is plentiful.

Second, consider a different way to look at the problem of figuring out what is \( \hat{\rho} \) and whether it is larger than \( \rho \). Recall that according to the parsimonious model, agents perceive that: \( z_t = \hat{\rho} z_{t-1} + \epsilon_t \). It then follows that, no matter what the signals are, the agent’s expectations at long and short horizons will be linked by: \( E_{i,t}(z_{t+T}) = \hat{\rho}^T E_{i,t}(z_t) \). Fortunately, the same survey that asks people what are their expectation for inflation over the next year, also asks for their expected inflation on average over the next 5 years. The ratio
of long-horizon and short-horizon expectations is:

\[
\sum_{h=1}^{H} \frac{\mathbb{E}_{i,t}(z_{t+h})}{H} = \frac{1}{H} \frac{1 - \hat{\rho}^H}{1 - \hat{\rho}}
\]  

(9)

which reveals what is the implicit \(\hat{\rho}\).

By taking the ratio of the impulse responses of long-horizon and short-horizon forecasts to the same shock, we can use the formula on the right-hand side of this equation to back out the \(\hat{\rho}\) in people’s minds. Intuitively, if that ratio is close to 1, it must be that people expect the long-run impact to be as large as the short-run one, and thus that the series is very persistent: a \(\hat{\rho}\) close to 1. Instead, if the ratio is small, then people think that the impact of the shocks on inflation dies off quickly so \(\hat{\rho}\) is close to 0. In the SPF data, this ratio reveals a \(\hat{\rho} = 0.81\). Using the actual inflation data and the estimates in figure 1, suggest a \(\rho = 0.26\), consistent with people over-extrapolating.

Following precisely the same procedure used to produce figure 1, estimate instead the impulse response of outcomes and forecasts for 5-year inflation and 5-year ahead forecasts. The results are in figure 2. Because annual inflation has a large transitory component, estimates of \(\rho\) based on figure 1 point to a low \(\rho\). But, when one looks at 5-years averages, as in figure 2, the permanent component stands out, and reveals that the right inference at longer horizons is that \(\rho\) is close to 1. More precisely, the largest autoregressive root of the inflation process is close to one, while the serial correlation at an annual frequency is close to zero. By comparison, the forecast estimates in figure 2 are always below the actual ones. While the forecasts decline towards zero, the outcomes do not. People behave as if inflation is a transitory process, in spite of the permanent component visible in outcomes. According to figure 2, we see people under-extrapolating.

There is a blunter way to state the story these estimates are telling. U.S. history teaches us that inflation can sometimes drift away, but only does so rarely. When asked in surveys about what they expect at long horizons, people at first respond too much to shocks, but then quickly revert to answering 2% no matter what the shock was. Is this over- or under-extrapolation?

For now, at least, the jury is still open on the direction of how \(\hat{R}(L)\) differs from \(R(L)\), that is on what is the best baseline choice of parameters for the parsimonious model of expectations.
5 Missing disagreement

Accusing a parsimonious model of missing some features misses the point of what parsimony is for. At the same time, for a model to clear the high bar that I set at the start—to become the Cobb-Douglas, the CRRA or the Calvo of expectations—then it should capture the central features that we have learned from surveys of expectations. The proposal by Angeletos, Huo and Sastry (2020) captured by equation (3) in section 2 does very well, but misses in one important dimension: disagreement and communication.

A large literature has studied the extent of disagreement in expectations, and how communication policies affect it. In theory, this has led to important insights on to the role of strategic complementarities and the effects of transparency (Morris and Shin, 2002, Haldane and McMahon, 2018). In policy, it has spurred insights on how to steer this disagreement and have a real effect on macroeconomic outcomes (Coibion, Gorodnichenko and Weber, 2019, Coibion et al., 2020). In the data, it has led to a focus on the second moment of expectations surveys (Mankiw, Reis and Wolfers, 2004, Dovern, Fritsche and Slacalek, 2012).

My reading of this literature is that a benchmark model should try to capture three important facts about disagreement on inflation:
1. Shocks, positive or negative, raise disagreement temporarily.

2. Policy communication lowers the disagreement that results from a shock.

3. Regime changes that raise transparency can permanently lower disagreement.

Within the framework of section 2 for a scalar fundamental, a measure of disagreement is the cross-sectional variance:

\[ V_t = \int \left( E_{i,t} (z_{t+1}) - \int E_{i,t} (z_{t+1}) dt \right)^2 di. \]  \hspace{1cm} (10)

Then, assuming an AR(1) with normal shocks for the fundamental \( z_t = \rho z_{t-1} + r \varepsilon_t \), and adding noisy normal signals \( z_t + \tau^{-1/2} u_{i,t} \), a few steps of algebra show that:

\[ V_t = \hat{\lambda}^2 V_{t-1} + (\hat{\rho} - \hat{\lambda})^2 \]  \hspace{1cm} (11)

where \( \hat{\lambda} \) is the root in \((0, \hat{\rho})\) of the quadratic: \( \hat{\lambda} + 1/\hat{\lambda} = \hat{\rho} + (1 + \tau)/\hat{\rho} \). Crucially, the shocks \( \varepsilon_t \) do not show up. This is a deterministic equation with a stable steady state. If disagreement starts at the steady, it will stay constant forever. It therefore follows that in the Angeletos, Huo and Sastry (2020) setup:

1. Disagreement does not respond to shocks.

2. Communication, understood as lowering \( r \) so that shocks are not as intense, has no effect on disagreement.

3. An increase in transparency, understood as more precise signal \( \tau \), raises disagreement as it lowers \( \lambda \).

In short, the model does not capture the endogeneity of disagreement in this literature.

6 Disagreeing constructively

A small modification of the model delivers an alternative parsimonious model that can capture disagreement. It adds one parameter \( \theta \), although since I had earlier removed another parameter, \( \hat{\tau} \), the model is arguably just as parsimonious. The modification works as follows: a fraction \( 1 - \theta \) of agents form expectations precisely according to equation (3). But, a fraction \( \theta \) instead happens to have full information, and so they know the
current state \( z_t \), making forecasts according to: \( \mathbb{E}_t(z_{t+1}) = \rho z_t \). Whether this fraction \( \theta \) is endogenous, or how it is drawn from the population, matters for dynamics. Different articles have explored how and why this matters, but in a simple parsimonious model \( \theta \) can just be taken as a given constant. Likewise, assuming that these agents have perfect information is just a simple way to capture heterogeneity in the population over their signal precision \( \tau \). Assuming different groups, each with a different \( \tau \), would simply be less parsimonious than assuming that a fraction has infinite precision, while the other fraction has a finite \( \tau \) precision.

A little bit of algebra shows that, in this model, the law of motion for disagreement is instead:

\[
(1 - \hat{\lambda})^2 V_t = \theta (1 - \theta)^2 \hat{\lambda}^2 \varepsilon_t^2 + (1 - \theta) (\hat{\rho} - \hat{\lambda})^2. \tag{12}
\]

Note that if \( \theta = 0 \), then equation (12) becomes equation (10).

In this parsimonious model, the under and over-reaction of expectations, the sluggishness in adjustment to shocks, and potentially the delayed overshooting, are all still present. But, disagreement is now endogenous and it has properties that fit the lessons from the data:

1. Disagreement varies over time and is affected by shocks, as it follows an AR(2) after a shock \( \varepsilon_t \).

2. Policy communication lowers disagreement, since a lower \( r \) lowers disagreement both on impact and in the steady state.

3. Transparency in the sense of a higher \( \theta \) lowers disagreement. The effect of a higher \( \tau \) on disagreement depends, as it changes both the reliance of agents on their signals, as well as the dispersion of these signals.

7 Conclusion

The literature on expectations in macroeconomics needs a simple canonical model, that can be used for teaching in core classes and as a benchmark assumption that replaces rational expectations. Angeletos, Huo and Sastry (2020) propose one such benchmark, with few parameters, that is easy to solve, and that can be incorporated in most dynamic macroeconomic models. Their model captures the under and over-reaction of forecasts, as well as the sluggish response of average expectations to shocks, that are the staple
results of this literature. It also can capture a candidate new fact on delayed overshooting of expectations through over-extrapolation. It fails to capture one important dimension of the literature, on disagreement and communication, but a modification of it can easily fix this problem. This modified imperfect expectations model meets all the criteria for the desired benchmark model. I hope teachers and researchers will use it, so that imperfect expectations soon become as routinely used as rational expectations have been so far.

References


