Macroeconomic Expectations and Noisy Memory

Yeji Sung

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Macroeconomic Expectations and Noisy Memory

Yeji Sung

A large empirical literature has documented that people often react too much to recent information compared to the rational benchmark. In this thesis, I propose an explanation for overreaction based on the idea of limited memory. Using information-theoretic constraints, I formalize that past knowledge is recalled with random errors (hence the "noisy memory"). Since forecasts are not accurately based on past knowledge, revising one's views more aggressively is optimal. While this mechanism explains over-reaction in general, I focus on specific applications in three chapters of this thesis. In the first two chapters, I explore how noisy memory impacts the learning of structural parameters. Specifically, I focus on learning about mean and variance of a stochastic process in each chapter. In the third chapter, I study how noisy memory interacts with conventional information frictions.

Table of Contents

Acknowledgments ix

Dedication x

Introduction or Preface 1

Chapter 1: Optimally Imprecise Memory and Biased Forecasts 4

1 Introduction 4

2 A Flexible Model of Imprecise Memory: A Simple Example 6

2.1 Feasible memory structures 8

2.2 The optimal memory structure 10

2.3 Implications for forecast dynamics and forecast errors 13

3 The Optimal Memory Structure when the State is Persistent 15

3.1 Implications of linear-Gaussian dynamics 20

3.2 The sufficiency of memory of a reduced cognitive state 22

3.3 A recursive formulation 26

3.4 Optimality of a unidimensional memory state 28

4 Features of the Model Solution 33

4.1 The case of a fixed per-period bound on mutual information 34

4.2 The case of a linear cost of information 40

4.3 Stationary fluctuations in the long run . . . 45

4.4 "Recency bias" in expectation formation . . . 49

5 Experimental Evidence . . . 51

6 Related Models . . . 56

6.1 Alternative Explanations for Over-Reaction . . . 56

6.2 Alternative Models of Imprecise Memory . . . 63

7 Conclusion . . . 68

8 Appendix . . . 72

8.1 Reduction of the General Forecasting Problem to Estimation of μ . . . 72

8.2 Bayesian Updating After the External State is Observed: A Kalman Filter . . . 75

8.3 Demonstration that an Optimal Memory Structure Records Information Only about the Reduced Cognitive State . . . 79

8.4 The Law of Motion for the Memory State and the Information Content of Memory . . . 88

8.5 Recursive Determination of the Optimal Memory Structure . . . 98

8.6 Numerical Solutions . . . 113

8.7 Predicted Values for the Quantitative Measures of Forecast Bias . . . 125

Chapter 2: Inflation Surprises and Perception of Inflation Risks . . . 133

1 Introduction . . . 133

2 Model . . . 135

2.1 Learning with Perfect Memory . . . 137

2.2 Learning with Noisy Memory . . . 138

2.3 Illustrating the Implications of Noisy Memory . . . 143

2.4 Testable predictions . . . 145
3 Empirical Evidence . . . 149
3.1 Data . . . 149
3.2 Estimates . . . 151
4 Conclusion . . . 153
5 Appendix . . . 156
5.1 Detailed Derivations . . . 156
5.2 Robustness Checks . . . 158

Chapter 3: Macroeconomic Expectations and Cognitive Noise . . . 160
1 Introduction . . . 160
2 A Model of Mental Representation . . . 165
2.1 The Forecasting Problem . . . 165
2.2 Mental Representation of External Information . . . 167
2.3 Mental Representation of Internal Information . . . 168
2.4 Forecasts Based on Represented Information . . . 170
2.5 The Nature of Information Frictions . . . 171
3 The Optimal Mental Representation . . . 172
3.1 The Optimization Problem . . . 172
3.2 Optimal Representation of Noisy News . . . 172
3.3 Optimal Representation of Noisy Memory . . . 174
4 Cognitive Noise and Biased Forecasts . . . 175
4.1 Forecasts Subject to Cognitive Constraints . . . 176

4.2 Biases in Survey Forecasts . . . 179

5 Extended Model . . . 182

5.1 Learning about the Long Run . . . 183

5.2 The Optimal Cognitive Process . . . 184

5.3 Perpetual Uncertainty about the Long Run . . . 188

6 Estimating the Extent of the Cognitive Constraints . . . 191

6.1 Data . . . 191

6.2 Estimation Strategy . . . 192

6.3 Estimation Results . . . 194

7 An Illustrative Macroeconomic Model . . . 203

7.1 Firms' Decision Problem . . . 203

7.2 Aggregate Economy . . . 205

7.3 Firms' Macroeconomic Expectations . . . 207

7.4 Expectation Formations and Inflation Dynamics . . . 208

7.5 Calibration . . . 210

7.6 Monetary Policy and Inflation Variability . . . 210

7.7 Efficient Inflation Targeting . . . 212

8 Conclusion . . . 216

9 Accompanying Tables and Figures . . . 218

10 Appendix . . . 221

10.1 Derivation of the Optimal Cognitive Process . . . 221

10.2 Optimal Covnitive Process When z_t is the Only State Variable . . . 225

10.3 Optimal Cognitive Process When (μ, z_t) is the State Vector . . . 227

10.4 Derivations of β_I and β_C (when the long-run mean is known) 237

10.5 Estimation 246

10.6 A Stationary Relationship 260

10.7 Monetary Model 268

References 281

List of Figures

1.1 The evolution of uncertainty about μ (for varying degress of $\bar{\lambda}$) 35

1.2 The optimal memory structure in the long run (for varying degress of $\bar{\lambda}$) . . 37

1.3 The evolution of uncertainty about μ (for varying degress of $\tilde{\theta}$) 41

1.4 The optimal memory structure in the long run (for varying degress of $\tilde{\theta}$) . . 42

1.5 The long-run uncertainty 43

1.6 Impulse responses 47

1.7 Two regression coefficients 54

1.8 The evolution of uncertainty about μ (when $\rho = 0$) 114

1.9 The evolution of uncertainty about μ (when $\rho > 0$) 115

1.10 The optimal policy function 117

1.11 The dynamics of scaled uncertainty and memory precision 118

1.12 The Bellman equation 120

1.13 The optimal policy function (for a sufficiently large $\tilde{\theta}$) 122

1.14 Impulse responses of the DM's estimate of μ for alternative degrees of persistence ρ of the external state process. 129

1.15 Impulse responses of the DM's one-quarter-ahead forecast of the external state for alternative degrees of persistence ρ of the external state process. . . 130

2.1 More Weight on Recent Surprises 146

2.2 Posterior Uncertainty about ω^{-1} . . . 147

2.3 Impulse Response Function of $E\left[\omega^{-1}|m_t, y_t\right]$ . . . 148

2.4 Scatter Plot of Baseline Regression . . . 152

2.5 Scatter Plot of Baseline Regression: Robustness . . . 158

3.1 Impulse response of forecasts . . . 178

3.2 β_C and β_I jointly identify the extent of cognitive noise . . . 183

3.3 Impulse-response functions when learning about the long run . . . 197

3.4 β_C and β_I when learning about the long run . . . 198

3.5 Estimated β_C and β_I . . . 199

3.6 Estimated Parameters . . . 200

3.7 Not-targeted moments (all macroeconomic variables) . . . 202

3.8 The effect of monetary policy . . . 214

3.9 Efficient policy . . . 215

List of Tables

2.1 Summary Statistics 154

2.2 Empirical Estimations: Inflation Surprises and Perceived Inflation Risk . . . 155

2.3 Empirical Estimation: Robusntess 159

3.1 Estimated regression coefficients 196

3.2 Model Fit 201

3.3 Estimated regression coefficients using OLS 218

3.4 Estimated parameters 219

3.5 Estimated parameters using Coibion and Gorodnichenko (2015) approach . 219

3.6 Model fit 220

3.7 Baseline Regression Coefficients 254

3.8 Regression Coefficients for Current-quarter Forecasts Only 255

3.9 Regression Coefficients for three-quarter-ahead Forecasts Only 256

3.10 Regression Coefficients Using All Sample Periods 257

3.11 Estimationg Using the Long-term CPI Forecasts 258

3.12 Data Generating Process: AR(1) process 259

Acknowledgements

I am indebted to my main advisor Michael Woodford for literally everything. The most powerful guidance I have received was to observe how Mike leads by example. I also thank my advisors, Jennifer La'O and Hassan Afrouzi, for always extending help when I needed it (and when I didn't even know I needed it). All my advisors are incredible role models, and I am forever grateful.

I have been enormously lucky to be a part of the Columbia community. I learned a lot from the macro group, and Macro Tuesday has been my favorite day of the week. Another equally tremendous experience at Columbia is the Cognitive Decision lab. It provided me with a safe environment to explore and mature, and I cannot thank Mark Dean enough for cultivating this culture. I am grateful to have met and interacted with Miguel Acosta, Motaz Al-Chanati, Paul Bouscasse, Juan Herreño, Vinayak Iyer, Paul Koh, Sun Lee, Wendy Morrison, Ricardo Pommer, Susie Scanlan, Maggi Shi, Edward Shore, and Sophie Zhou.

I also want to thank mentors from a broader macro community. Jae Won Lee and Jay Hong are my first mentors from Seoul National University who taught me so many things when I knew so little. This dissertation has also benefitted from conversations with economists at the Federal Reserve during my two summer visits, 2021 in D.C. and 2022 in St. Louis.

Finally, my deepest gratitude goes to my parents and sister — Soohyun Kwon, Nakil Sung, Yewon Sung — and Dan Son. They made this dissertation possible.

To my family

Preface

Despite the popularity of the rational expectations hypothesis (REH), people consistently hold expectations that are inconsistent with the optimal Bayesian inference. Many patterns of such behavior are documented by long and extensive empirical literature that is almost as old as the history of modern macroeconomics. In particular, it is often observed that people react too much to new information compared to what they ought to, given the knowledge they ought to have. This notion of over-reaction has been a popular explanation for several perplexing features of asset price dynamics, such as the high premium of holding stocks that cannot be readily justified by conventional models. While many explanations have been proposed to explain why people react too much, it turned out that the pattern of observations is far more prevalent than what these candidate explanations can make sense of.

In this thesis, I propose another explanation for overreaction by formalizing the idea that human judgments are based on limited memory. The constraint is a limit on the complexity of memory measured using Shannon's mutual information, as in models of rational inattention by Sims (2003). The difference is that retrieving past cognitive states is not costless. The consequence of this memory constraint is that past knowledge is recalled with random errors (hence the "noisy memory"). I optimally derive the process in which knowledge is accumulated with random noise. The presence of such information friction implies that learning is slower and less accurate compared to the perfect memory bench-

mark. Since prior knowledge resolves less uncertainty about the state of the economy, it is optimal to revise one's views more aggressively. In the three chapters of the thesis, I explore specific applications of noisy memory.

The first chapter, written with Rava A. da Silveira and Michael Woodford, introduces the formulation of memory constraints used throughout the thesis. We focus on a particular example of a decision-maker learning about the mean of a stochastic process. The model implies that forecasts exhibit idiosyncratic random variation and that even average beliefs also differ from rational-expectations beliefs. The bias in forecasts fluctuates forever with a variance that does not fall to zero, even in the long run. In addition, more recent news will be given disproportionate weight in forecasts. We solve the model under various assumptions (such as the degree of persistence of the variable to be forecasted and the horizon over which it must be forecasted) and examine how the nature of forecast biases depends on these parameters.

The second chapter, written with Miguel Acosta, documents that people become more uncertain about future inflation when current inflation deviates from their expectations. This behavior is a feature of professional forecasts, even when inflation is low and stable. Typical models of expectation formation leave little room for this link between inflation surprises and perceived inflation risk, especially when inflation has long been stable. We propose limited memory as an explanation for this link. When people learn about the inflation process from their experience, they perceive inflation to be more variable when faced with large variations in inflation. This intuition has implications for the price determination of financial contracts that compensate for the variation of the underlying assets.

The third chapter incorporates noisy memory into conventional models of information frictions to explain the puzzling patterns observed in survey forecasts. One conventional motivation for information frictions is that it is costly to process information. I propose a model in which one's past knowledge (stored in memory) is also costly to process. The model is consistent with survey-forecast patterns and offers an estimation strategy

to identify the extent of information frictions. I then explore the macroeconomic implications of these frictions. The proposed model suggests that inflation expectations are not well anchored, making it more challenging to stabilize inflation than under conventional information-friction models.

Chapter 1: Optimally Imprecise Memory and Biased Forecasts

with Rava Azeredo da Silveira[1] and Michael Woodford[2]

1 Introduction

The hypothesis of rational expectations (RE) proposes that decisions are based on expectations that make use of all available information in an optimal way: that is, those that would be derived by correct Bayesian inference from an objectively correct prior and the data that has been observed to that date. Yet both in surveys of individual forecasts of macroeconomic and financial variables and in forecasts elicited in experimental settings, beliefs are more heterogeneous than this hypothesis should allow, and forecast errors are predictable on the basis of variables observable by the forecasters, contrary to this hypothesis. In particular, a number of studies have argued that forecasts typically over-react to new realizations of the variable being forecasted. (See Bordalo *et al.,* 2020, and Afrouzi *et al.,* 2020, for recent examples with extensive references to prior literature.)

Here we offer an explanation for the pervasiveness of over-reaction, that depends neither on an assumption that people follow arbitrary (and distinctly sub-optimal) heuristics, or that their forecasts make sense only under an incorrectly specified statistical model. We propose a theory in which a decision maker's forecasts (or more generally, actions with consequences that depend on the future realization of some variable) can be based both on currently observable information and an imperfect memory of past observations. Subject to this constraint on the information that the decision rule can use, we assume that their decision rule is optimal. Moreover, rather than making an arbitrary assumption about the kind of statistics about past experience that can be recalled with greater or lesser precision,

[1]ENS and University of Basel
[2]Columbia University

we allow the memory structure to be specified in a very flexible way, and assume that it is optimized for the particular decision problem, subject only to a constraint on the overall complexity of the information that can be stored in (and retrieved from) memory — or more generally, subject to a cost of using a more complex memory structure.

In the limiting case in which the cost of memory complexity is assumed to be negligible, the predictions of our model coincide with those of the rational expectations hypothesis. But when the cost is larger (or the constraint on memory complexity is tighter), our model predicts that forecasts should be both heterogeneous (even in the case of forecasters who observe identical data) and systematically biased. Moreover, the predicted biases include the type of over-reaction to news documented in surveys of forecasts of macroeconomic and financial time series by Bordalo *et al.* (2020) and in laboratory forecasting experiments by Afrouzi *et al.* (2020). And unlike the theory of "natural expectations"l of Fuster *et al.* (2010, 2011), our model predicts that over-reaction to news will be most severe in the case of time series exhibiting little serial correlation.

In seeking to endogenize the information content of the noisy cognitive state on the basis of which people must act, our theory is in the spirit of Sims's (2003) theory of "rational inattention"; and indeed, we follow Sims in modeling the complexity constraint using information theory. There is nonetheless an important difference between our theory and that of Sims (2003). Sims assumes a constraint on the precision with which new observations of the world can reflect any current or past conditions outside the head of the decision maker, but assumes perfectly precise memory of all of the decision maker's own past cognitive states, and also assumes that past external states can be observed at any time with the same precision as current conditions. We instead assume (for the sake of simplicity) that the current external state can be observed with perfect precision, but that memory of past cognitive states is subject to an information constraint; and we further assume that the decision maker has no access to external states that occurred in the past, except through (information-constrained) access to her own memory of those past states.

These differences are crucial for the ability of our model to explain over-reaction to news.[3]

In section 1, we present the assumptions of our model of endogenously imprecise memory, and illustrate its consequences for a simple example in which the variable to be forecasted is i.i.d. Section 2 then offers a more general characterization of the optimal memory structure in our model, showing in particular that even when the variable to be forecasted is serially correlated, it is optimal under our assumptions for the memory state at each point in time to be represented by a single real number, a random variable the mean of which depends on the entire sequence of previous observations. Section 3 illustrates the model's implications, discussing quantitative aspects of numerical solutions of the model for particular parameter values. We emphasize the failure of beliefs ever to converge to those associated with a rational expectations equilibrium, and show that instead, there are perpetual stationary fluctuations in subjective beliefs similar (though not identical) to those predicted by models of "constant-gain learning" (Evans and Honkapohja, 2001). Finally, section 4 compares the quantitative predictions of the model to the reported expectations of subjects in the laboratory experiment of Afrouzi *et al.* (2020), showing not only that the model can produce over-reaction to news, but that it can be parameterized so as to predict roughly the degree of over-reaction that is observed. Section 5 compares our model with alternative explanations for over-reaction of expectations, some of which are based on alternative models of imperfect memory, and section 6 concludes.

2 A Flexible Model of Imprecise Memory: A Simple Example

Here we precisely specify the constraint on the precision of memory that we propose, and illustrate the kind of conclusions that follow from it by first discussing a simple case, in which the state variable to be forecasted is an i.i.d. random variable. The problem of

[3]Other recent papers that explore the consequences of assuming that memory allows only a noisy recollection of past observations include Afrouzi *et al.* (2020) and Neligh (2022). While these authors also assume that some aspects of memory structure are optimized for a particular decision problem, the classes of memory structures that they consider are different than the one that we analyze here. See section 5.2 for further discussion.

the decision maker [DM] is generalized in the following section.

Suppose that the variable y_t is an independent draw each period from a Gaussian distribution, $y_t \sim N(\mu, \sigma_y^2)$, and that the DM's problem at each time t is to produce a forecast z_t of the value of y_{t+h} — that is, the value of the external state that will be observed h periods later (for some $h \geq 1$). The forecast z_t is produced after observing the value of y_t. If we suppose that the DM's loss from making an inaccurate forecast is proportional to the squared error of the forecast, then an optimal forecasting rule (subject to the memory constraints to be specified below) will be one that minimizes the expected value of the discounted quadratic loss function

$$\sum_{t=0}^{\infty} \beta^t (z_t - y_{t+h})^2, \tag{2.1}$$

where $0 < \beta < 1$ is the DM's discount factor.

Given that future realizations of the state are completely independent of anything observed in the past, it is obvious that if the parameters of the distribution from which y_t is drawn are known (that is, if the DM's decision rule can be designed using the values of these parameters), then the optimal forecast each period will simply be $z_t = \mu$, the unconditional expected value of y_{t+h}. We assume, however, that the DM's decision rule must be chosen without knowledge of the value of μ; instead, the decision is optimized for a prior over the possible values of μ, $\mu \sim N(0, \Omega)$, for some $\Omega > 0$. In this case, an optimal decision rule will seek to estimate the value of μ (and hence the minimum-mean-squared-error [MMSE] forecast) from observations of the state that have been made up to time t.

We simplify the discussion by supposing that the value of σ_y^2 is known (can be used in specifying the DM's decision rule); this makes it straightforward to say how much can be inferred about the value of μ from an observation of the state y_t. In the case of perfect memory, so that the DM's forecast z_t can be a function of the complete sequence of observations $(y_0, \ldots, y_t)$ from some initial period zero onwards, the computation of the MMSE

estimate of μ is a standard Kalman-filtering problem. Posterior beliefs after y_{t-1} has been observed are of the form $\mu \sim N(\hat{\mu}_{t-1}, \hat{\sigma}^2_{t-1})$, where the mean and variance of this Gaussian distribution are to be calculated. It then follows that after the next observation y_t, the new posterior will be of the same form, with mean and variance given by the recursions

$$\hat{\mu}_t = \hat{\mu}_{t-1} + \gamma_t(y_t - \hat{\mu}_{t-1}),$$

$$\hat{\sigma}^2_t = \frac{\hat{\sigma}^2_{t-1}\sigma^2_y}{\hat{\sigma}^2_{t-1} + \sigma^2_y},$$

where the "Kalman gain"

$$\gamma_t = \frac{\hat{\sigma}^2_{t-1}}{\hat{\sigma}^2_{t-1} + \sigma^2_y}$$

is a factor between 0 and 1. These equations can be solved recursively to determine $\hat{\mu}_t$ and $\hat{\sigma}^2_t$ for all t (given a sequence of realizations of the state), starting from initial conditions $\hat{\mu}_{-1} = 0$, $\hat{\sigma}^2_{-1} = \Omega$.

These equations imply that the precision $\hat{\sigma}_t^{-2}$ grows linearly with the number of observations, and hence that $\hat{\sigma}^2_t \to 0$ as $t \to \infty$, regardless of the sequence of observations. Eventually the correct value of μ is learned to arbitrary precision, and new observations cease to affect the estimate of μ ($\gamma_t \to 0$), and consequently cease to affect the DM's forecast. Thus forecasts are eventually the same as under an assumption of (full-information) rational expectations. We wish to examine how these conclusions change in the case of imperfect memory.

2.1 Feasible memory structures

We assume that the memory carried into each period $t \geq 0$ can be summarized by a vector m_t of dimension d_t; the action chosen in period t (i.e., the choice of $\hat{\mu}_t$) must be a function of the cognitive state specified by $s_t = (m_t, y_t)$. The dimension of the memory state is assumed only to be finite, and is not required to be the same for all t. (The case of *perfect memory* can be accommodated by our notation, by assuming that $d_t = t$, and

that the elements of the vector m_t correspond to the values $(y_0, y_1, \ldots, y_{t-1})$.) We assume that current external state y_t is perfectly observable,[4] but that behavior can depend on past states only to the extent that memory provides information about them.

We further suppose that the memory state evolves according to a linear law of motion of the form

$$m_{t+1} \;=\; \Lambda_t s_t \;+\; \omega_{t+1}, \qquad \omega_{t+1} \;\sim\; N(0,\, \Sigma_{\omega,t+1}) \tag{2.2}$$

starting from an initial condition of dimension $d_0 = 0$ (that is, s_0 consists only of y_0). However, the dimension d_{t+1} of the memory that is stored, and the elements of the matrices $\Lambda_t, \Sigma_{\omega,t+1}$ are allowed to be arbitrary; we require only that $\Sigma_{\omega,t+1}$ must be positive semi-definite (though it need not be of full rank).

For example, one type of memory structure that this formalism allows us to consider is an "episodic" memory of the kind assumed by Neligh (2022).[5] In this case, $d_t = t$, and there is an element of m_t corresponding to each of the past observations y_τ for $0 \leq \tau \leq t-1$ (generalizing the case of perfect memory just discussed). The memory of y_τ at some later time t is given by $m_{\tau+1,t} = y_\tau + u_{\tau+1,t}$, where $u_{\tau+1,t}$ is a Gaussian noise term, independent of the value of y_τ, and with a variance that is necessarily non-decreasing in t. This can be represented by letting $d_t = t$, Λ_t be the identity matrix of dimension $t+1$, and $\Sigma_{\omega,t+1}$ a diagonal matrix of dimension $n+1$ (with non-negative elements, but not necessarily of full rank).

Another type of memory that we can consider is one in which only the n most recent past observations of the external state can be recalled, though these are recalled with perfect precision. The requirement that forecasts be functions of the cognitive state would then require them to be functions of $(y_t, y_{t-1}, \ldots, y_{t-n})$ for some finite n, as under the hypothesis of "natural expectations" proposed by Fuster, Hébert, and Laibson (2011). This

[4] The case in which the current state is observable only imprecisely is discussed in Sung (2022).

[5] Note however that Neligh's model is not a special case of ours, because in addition to restricting attention to a more special class of memory structures, he assumes a different cost function for precision than the one we propose below.

case would correspond to a specification in which $d_t = n$ for all t; Λ_t is an $n \times (n+1)$ matrix, the right $n \times n$ block of which is an identity matrix, and the first column of which consists entirely of zeroes; and $\Sigma_{\omega,t+1} = 0$. Our formalism is much more flexible than either of these cases, however, and neither of those specifications turns out to be optimal.

We limit the precision of memory by further assuming that there is a cost of storing and/or accessing the memory state m_{t+1}, that is an increasing function of the Shannon mutual information between the memory state m_{t+1} and the cognitive state s_t about which it provides information.[6] In this section, we assume that there is a finite upper bound $\bar{I}$ on the feasible rate of information transmission: thus feasible memory structures must satisfy the constraint $I_t \leq \bar{I}$, where I_t is the mutual information between s_t and m_{t+1}.[7] Subject to this constraint on feasible memory structures, both the memory structure and the decision rule (specifying z_t as a function of the cognitive state s_t) each period are chosen so as to achieve the minimum possible expected value of (2.4).

2.2 The optimal memory structure

Here we sketch the implications of this model of noisy memory for the simple forecasting problem introduced above. (A more complete presentation of the calculations is offered below, where we also discuss a more general problem.) If we introduce the notation

$$\mu \,|s_t \;\sim\; N(\hat{\mu}_t,\, \hat{\sigma}_t^2) \tag{2.3}$$

for the posterior distribution for μ conditional on the cognitive state s_t, we observe that the DM's optimal decision rule will be $z_t = \hat{\mu}_t$ each period, and that the minimum achievable

[6] Mutual information is a non-negative scalar quantity that can be defined for any joint distribution for (s_t, m_{t+1}), that measures the degree to which the realized value of either random variable provides information about the value of the other (Cover and Thomas, 2006). This measure is used to determine the relative cost of different information structures in the rational inattention theory of Sims (2003); properties of this measure as an information cost function are discussed in Caplin, Dean and Leahy (2019).

[7] In section 2, we generalize this assumption to allow I_t to be increased at some positive marginal cost.

value for the loss function (2.4), given the memory structure, will be

$$\sum_{t=0}^{\infty} \beta^t [\hat{\sigma}_t^2 + \sigma_y^2].$$

It follows that the optimal memory structure will be the one that minimizes the implied value of

$$\sum_{t=0}^{\infty} \beta^t \hat{\sigma}_t^2, \tag{2.4}$$

When $\{y_t\}$ is an i.i.d. random variable, the only possible relevance of memory for decisions in periods $t+1$ or later is the evidence that memory can provide about the value of the parameter μ. Hence the only aspect of the cognitive state s_t that is worth remembering later is what was known then about the value of μ, which is to say, the parameters of the distribution (2.3). Given the linear-Gaussian dynamics in our model, one can show that $\hat{\sigma}_t^2$ is independent of the history of realizations of the external state, and hence the same in all possible cognitive states s_t.[8] Thus the scalar quantity $\hat{\mu}_t$ is the only aspect of the cognitive state that is worth remembering.

Since a memory state m_{t+1} that was informative about any other aspect of s_t would increase the value of I_t without increasing the information provided about the value of μ, an optimal memory structure will make the distribution of the random variable $m_{t+1}\,|s_t$ a function only of $\hat{\mu}_t$. Under the assumption of linear-Gaussian dynamics (2.2), we must therefore be able to write

$$m_{t+1} \;=\; \Lambda_t\, \hat{\mu}_t \;+\; \omega_{t+1}, \tag{2.5}$$

where Λ_t is now a column vector rather than a matrix.

Given the memory state m_t that can retrieved in any period, the implied posterior

[8] It depends on t, which is to say the number of observations that have occurred; but this is assumed to be available as an input to the decision rule, rather than something that has to be remembered using costly memory.

distribution for the parameter μ will be a Gaussian distribution,

$$\mu \,|m_t \;\sim\; N(\bar{m}_t,\, \Sigma_t).$$

(When memory is imperfect, however, we can no longer identify $(\bar{m}_t,\, \Sigma_t)$ with $(\hat{\mu}_{t-1},\, \hat{\sigma}^2_{t-1})$.) After the value of y_t is observed, these beliefs are updated to a posterior of the form (2.3), where

$$\hat{\mu}_t \;=\; \bar{m}_t \,+\, \gamma_t\,(y_t - \bar{m}_t), \tag{2.6}$$

using the notation

$$\gamma_t \;=\; \frac{\Sigma_t}{\Sigma_t + \sigma^2_y} \tag{2.7}$$

for the Kalman gain, and

$$\hat{\sigma}^2_t \;=\; \frac{\Sigma_t \sigma^2_y}{\Sigma_t + \sigma^2_y}. \tag{2.8}$$

Our linear-Gaussian framework further implies that $\bar{m}_t$ must be a linear function of m_t, while Σ_t is independent of m_t. It then follows from (2.5) that we can write

$$\bar{m}_{t+1} \;=\; \lambda_t\,\hat{\mu}_t \,+\, \bar{\omega}_{t+1}, \tag{2.9}$$

where now λ_t is a scalar, and $\bar{\omega}_{t+1} \;\sim\; N(0,\, \sigma^2_{\bar{\omega},t+1})$ a scalar random variable. We can further show that for any feasible memory structure, λ_t must be a quantity no less than zero and less than 1, and that

$$\sigma^2_{\bar{\omega},t+1} \;=\; \lambda_t(1-\lambda_t)\,\mathrm{var}[\hat{\mu}_t] \;=\; \lambda_t(1-\lambda_t)\,[\Omega - \hat{\sigma}^2_t].$$

Thus the law of motion (2.9) is fully specified by choosing a value for λ_t.

The only information about s_t contained in m_{t+1} must be the information about the value of $\hat{\mu}_t$ provided by the value of $\bar{m}_{t+1}$; hence under an optimal information structure, the value of I_t will be the mutual information between the random variables $\hat{\mu}_t$ and $\bar{m}_{t+1}$.

It follows from (2.9) that this is equal to $-(1/2)\ln(1-\lambda_t)$, an increasing function of λ_t. Thus the constraint $I_t \leq \bar{I}$ can alternatively be expressed as a constraint of the form $\lambda_t \leq \bar{\lambda}$, where $0 < \bar{\lambda} < 1$. (The limiting case in which $\bar{\lambda} = 1$ corresponds to no upper bound on the mutual information, and hence perfect memory.)

We can further show that the uncertainty about the value of μ in all periods $\tau > t$ is minimized (and hence the loss function (2.4) is minimized) by setting λ_t as large as possible, consistent with the constraint. Hence in each period the upper bound constraint will bind, and the optimal memory structure will be the one in which $\lambda_t = \bar{\lambda}$ each period. The law of motion (2.9) can accordingly be written

$$\bar{m}_{t+1} \;=\; \bar{\lambda}\,\hat{\mu}_t \;+\; \bar{\omega}_{t+1}, \tag{2.10}$$

and the associated posterior variance will equal

$$\Sigma_{t+1} \;=\; \Omega - \mathrm{var}(\bar{m}_{t+1}) \;=\; (1-\bar{\lambda})\Omega \;+\; \bar{\lambda}\hat{\sigma}_t^2. \tag{2.11}$$

Equations (2.6)–(2.8) and (2.10)–(2.11) then constitute a complete system of equations to recursively determine the evolution of the variables $\{\bar{m}_t, \Sigma_t, \hat{\mu}_t, \hat{\sigma}_t^2\}$ for all $t \geq 0$ given the sequence of observations $\{y_t\}$, starting from initial conditions $\bar{m}_0 = 0, \Sigma_0 = \Omega$ corresponding to the prior. (This generalizes the recursive system given above for the case of perfect memory.)

2.3 Implications for forecast dynamics and forecast errors

In the simple problem considered here, the optimal forecast each period is given by $z_t = \hat{\mu}_t$; the predictable part of the forecast error (if any) will simply be the predictable error (if any) in $\hat{\mu}$ as an estimate of μ; and the mean squared error of the forecast will equal $\hat{\sigma}_t^2 + \sigma_y^2$, where $\hat{\sigma}_t^2$ is the mean squared error of the estimate of μ. Thus we need only analyze the dynamics of the estimate $\hat{\mu}_t$ and the estimation error that this reflects.

In the perfect-memory case, the recursive system of equations presented above imply that

$$\hat{\mu}_t \;=\; \gamma_t \sum_{\tau=0}^{t} y_\tau,$$

so that each observation up through date t has an equal effect on the estimate (there are no "order effects"), and the optimal estimate is a positive multiple of the mean of the observed values $\{y_\tau\}$. The multiplicative factor $k_t = (t+1)\gamma_t$ is less than 1,[9] but converges to 1 as t becomes large (and $\gamma_t \to 0$).

In the noisy-memory ($\bar{\lambda} < 1$) case, instead, the solution for $\hat{\mu}_t$ is different in three important respects. First, the Kalman gain γ_t (which is again the weight on the current observation y_t) does not converge to 0 as t becomes large, but instead converges to a long-run value $\bar{\gamma}$ between 0 and 1. This is because the dynamics of the sequence $\{\hat{\sigma}_t^2\}$ implied by equations (2.8) and (2.11) imply that $\hat{\sigma}_t^2$ converges to a positive long-run value,[10] so that (2.7) then implies that γ_t converges to a positive value less than 1. Second, the weights on the different observations $\{y_\tau\}$ are not equal; instead the effect of a given observation on the estimate is smaller, the more distant the observation in the past (a "recency effect"). And third, instead of $\hat{\mu}_t$ being a deterministic function of the $\{y_\tau\}$, the estimate is also affected by the sequence of memory noise terms $\{\bar{\omega}_\tau\}$.

Specifically, one can write

$$\hat{\mu}_t \;=\; \sum_{j=0}^{t} \alpha_{j,t} y_{t-j} \;+\; \sum_{j=0}^{t} \beta_{j,t} \bar{\omega}_{t-j}, \tag{2.12}$$

where the coefficients are given by

$$\alpha_{j,t} \;=\; \bar{\lambda}^j \gamma_{t-j} \Pi_{i=1}^{j} (1 - \gamma_{t-j+i}), \tag{2.13}$$

$$\beta_{j,t} \;=\; \bar{\lambda}^j \Pi_{i=0}^{j} (1 - \gamma_{t-j+i})$$

[9] This reflects shrinkage of the Bayesian estimate of μ toward the prior mean of zero.

[10] See Figure 1 below for numerical examples, and Appendix F.4 for further analytical discussion.

for all $j \geq 0$.[11] In the limit as t becomes large, the coefficients converge:

$$\alpha_{j,t} \;\rightarrow\; \alpha_j \;\equiv\; \bar{\gamma}(\bar{\lambda}(1-\bar{\gamma}))^j,$$

$$\beta_{j,t} \;\rightarrow\; \beta_j \;\equiv\; \bar{\lambda}^j(1-\bar{\gamma})^{j+1}.$$

Thus in the large-t limit, the estimate $\hat{\mu}_t$ comes to equal a positive multiple of an exponentially-weighted moving average of past observations $\{y_\tau\}$, plus a serially-correlated noise term. Because of the exponentially decreasing weights, the term $\sum_j \alpha_j y_{t-j}$ continues to fluctuate randomly in response to the randomness in recent observations, rather than converging to the true value of μ with probability 1 (as in the perfect-memory case). Because $\sum_j \alpha_j < 1$, this term is also on average closer to 0 than is the true value of μ: the shrinkage toward the prior mean is not eliminated even as $t \rightarrow \infty$. And because the β_j are positive, the term $\sum_j \beta_j \bar{\omega}_{t-j}$ is an additional source of random variation in the estimate (and hence in the DM's forecast), independent of the sequence of observations $\{y_\tau\}$.

Because the limiting coefficients α_j are positive, the estimate $\hat{\mu}_t$ (and hence the DM's forecast z_t) continues to be influenced by recent observations y_{t-j} even when t is large — unlike the rational-expectations forecast, $z_t = \mu$. Thus the DM's forecast is predicted to "over-react" to news about recent observations.[12] Though the calculations required are more complex, we obtain qualitatively similar conclusions in the case that the DM forecasts a serially correlated variable, as we show next.

3 The Optimal Memory Structure when the State is Persistent

We now consider a more general class of linear-quadratic decision problems, allowing both for simultaneous forecasting of many different horizons, and for persistent dynamics in the state $\{y_t\}$ that is to be forecasted. We allow the state y_t to follow a stationary AR(1)

[11] When $j = 0$ in (2.13), we define the product with no factors to equal 1.

[12] We compare the predictions of our model to the measures of over-reaction reported by Afrouzi *et al.* (2020) in section 4.

process. We write its law of motion as

$$y_t \;=\; \mu \;+\; \rho(y_{t-1} - \mu) \;+\; \epsilon_{yt}, \tag{3.14}$$

where μ is again the mean, ρ is the coefficient of serial correlation (with $|\rho| < 1$), and $\{\epsilon_{yt}\}$ is an i.i.d. sequence, drawn each period from a Gaussian distribution $N(0,\, \sigma_\epsilon^2)$. The variance of the external state (conditional on the value of μ and the other parameters) will therefore equal $\sigma_y^2 \equiv \sigma_\epsilon^2/(1-\rho^2)$.

The DM's problem is to produce each period a vector of forecasts z_t, so as to minimize the expected value of a discounted quadratic loss function

$$\mathrm{E}\sum_{t=0}^{\infty} \beta^t (z_t - \tilde{z}_t)' W (z_t - \tilde{z}_t), \tag{3.15}$$

where W is a positive definite matrix specifying the relative importance of accuracy of the different dimensions of the vector of forecasts, and the eventual outcomes that the DM seeks to forecast are functions of the future evolution of the external state,[13]

$$\tilde{z}_t \;\equiv\; \sum_{j=0}^{\infty} A_j y_{t+j},$$

where the coefficients $\{A_j\}$ satisfy $\sum_j |A_j| < \infty$. (We again assume that $0 < \beta < 1$.) This formalism allows us to assume that the DM may produce forecasts about the future state at multiple horizons (as is typically true in surveys of forecasters, and also in the experiment of Afrouzi *et al.,* 2020). It also allows us to treat cases in which the DM may choose a vector of actions, the rewards from which are a quadratic function of the action vector and the external state in various periods; the problem of action choice to maximize expected reward in such a case is equivalent to a problem of minimizing a quadratic function of the

[13]Note that the variables denoted $\tilde{z}_t$ are not quantities the value of which is determined at time t; the subscript t is used to identify the time at which the DM must produce a forecast of the quantity, not the time at which the outcome will be realized. Thus the best possible forecast of $\tilde{z}_t$ at time t, even with full information, would be given by $\mathrm{E}_t \tilde{z}_t$, which will generally not be the same as the realized value $\tilde{z}_t$.

DM's error in forecasting certain linear combinations of the value of the external state at various horizons.[14]

To simplify our discussion, we continue to assume that the second moments of the stochastic process for the external state are known (more precisely, that the DM's decision rule can be optimized for particular values of these parameters, that are assumed to be the correct ones), while the first moment is not, so that the DM's decision rule must respond adaptively to evidence about the unknown mean value provided by the DM's observations of the state. Thus the values of the parameters ρ and σ_ϵ^2 are assumed to be known, while μ is not; the parameter μ is again assumed to be drawn from a prior distribution $\mu \sim N(0, \Omega)$. Conditional on the value of μ, the initial lagged state y_{-1} is assumed to be drawn from the prior distribution $N(\mu, \sigma_y^2)$, the ergodic distribution for the external state given a value for μ. When we consider the optimality of a possible decision rule for the DM, we integrate over this prior distribution of possible values for μ and y_{-1}, assuming that the decision rule must operate in the same way regardless of which values happen to be true in a given environment.

In any problem of this form (regardless of the assumed memory limitations), the DM's problem can equivalently be formulated as one of simply choosing an estimate $\hat{\mu}_t$ of the unknown mean μ at each date t, based on the information available at the time that z_t must be chosen. It follows from the law of motion (3.14) that

$$\mathrm{E}_t \tilde{z}_t \;=\; \sum_{j=0}^{\infty} A_j[\mu + \rho^j(y_t - \mu)],$$

where we use the notation $\mathrm{E}_t[\cdot]$ for the expected value conditional on the true state at time t, i.e., the value of μ and the history of realizations $(y_0, \ldots, y_t)$, even though not all of this information is available to the DM. Conditioning instead on the coarser information set

[14]For example, in a standard consumption-smoothing problem with quadratic consumption utility, the DM's level of expected utility depends on the accuracy with which "permanent income" is estimated at each point in time. This requires the DM to forecast a single variable $\tilde{z}_t$, for which the coefficient A_j is proportional to β^j for all $j \geq 0$.

that represents the DM's cognitive state at time t (and noting that this includes precise awareness of the value of y_t), we similarly find that the optimal estimate of $\tilde{z}_t$ will be given by

$$z_t \;=\; \sum_{j=0}^{\infty} A_j[\hat{\mu}_t + \rho^j(y_t - \hat{\mu}_t)], \tag{3.16}$$

where we again use the notation (2.3).

We show in the appendix that the DM's expected loss cannot be reduced by restricting attention to a class of decision rules of the form (3.16), under different possible assumptions about how the estimate $\hat{\mu}_t$ is formed.[15] In the case of any forecasting rule of that kind, the loss function (3.15) is equal to

$$\alpha \cdot \sum_{t=0}^{\infty} \beta^t MSE_t \tag{3.17}$$

plus a term that is independent of the DM's forecasts, where

$$MSE_t \;\equiv\; \mathrm{E}[(\hat{\mu}_t - \mu)^2]$$

is the mean squared error in estimating μ, and $\alpha > 0$ is a constant that depends on the coefficients $\{A_j\}$ and W. Thus one can equivalently formulate the DM's problem as one of optimal choice of an estimate $\hat{\mu}_t$ each period, so as to minimize MSE_t.

Feasible memory structures are again assumed to be described by linear-Gaussian dynamics of the kind specified in section 1.1. However, rather than assuming that there must be a fixed upper bound $\bar{I}$ on the mutual information I_t, we can assume more generally that there is a cost $c(I_t)$ of storing and/or accessing the memory state m_{t+1}, where $c(I)$ is an increasing and (at least weakly) convex function.[16]

The cost $c(I_t)$ can equivalently be viewed as either a cost of storing a memory record

[15] See Appendix A for details of the argument.

[16] The case of a fixed upper bound on the mutual information, considered above, can be nested as a special case of this model, in which $c(I) = 0$ for all $I \leq \bar{I}$, while the function is equal to $+\infty$ in the case of any $I > \bar{I}$.

with information content I_t (that is then available with perfect precision in period $t+1$), or a cost of retrieving a signal from memory with information content I_t in period $t+1$ (while the memory stored in period t is taken to have been a perfect record of the period t cognitive state). These two formulations are identical, given that we assume that only the signal m_{t+1} that is retrieved in period $t+1$ can be stored for future use; thus only the fidelity with which the retrieved memory m_{t+1} reproduces the cognitive state s_t matters. Under the retrieval-cost interpretation, however, our model remains importantly different from the one proposed by Afrouzi *et al.* (2020), in which memory contains a perfect record of all past observations, but there is a cost of retrieving a precise signal about the contents of memory for use in a decision. That model assumes that past observations can be stored indefinitely with perfect precision, with a limit on the precision of recall becoming relevant only when memory must be consulted; this means that it does not predict "recency bias" as ours does.[17]

The memory structure each period, together with the rule for choosing an estimate $\hat{\mu}_t$ as a function of each period's cognitive state, are then assumed to be chosen so as to minimize total discounted costs

$$\sum_{t=0}^{\infty} \beta^t \left[\alpha \cdot MSE_t + c(I_t),\right] \tag{3.18}$$

taking into account both the cost of less accurate forecasts (3.17) and the cost of greater memory precision. Note that no expectation is needed in this objective, since both MSE_t and I_t are functions of the entire joint probability distribution of possible values for $\mu, m_t, y_t, \hat{\mu}_t$ and m_{t+1}. We turn now to a general characterization of the solution to this dynamic opti-

[17] See the discussion in sections 3.4 and 5.2.2. The model of Afrouzi *et al.* also assumes that information that is retrieved from memory (at a cost) for use in a decision at time t has no consequences for the information that will be available at later times; the perfectly accurate record of all past observations continues to contain the same information regardless of what is retrieved at time t, while the information retrieved (added to "working memory") at time t is not available at any later time. This makes the problem of optimal selection of the information to be retrieved at any time t a (relatively simple) static problem in their model, whereas it is a dynamic problem in the model proposed here, since in our model, information not remembered at time t cannot (at any cost) be retrieved in any later period.

mization problem.

3.1 Implications of linear-Gaussian dynamics

For any memory structure in the class specified in section 1.1, the posterior distribution over possible values of $(\mu, y_{-1}, y_0, \ldots, y_{t-1})$ implied by memory state m_t will be a multivariate Gaussian distribution. It is thus fully characterized by specifying a finite set of first and second moments of the posterior associated with the memory state. Moreover, the particular memory state m_t at a given date t can be identified by the associated vector of first moments. For the second moments of the posterior are the same for all possible memory states at any time t: they depend on the matrices $\{\Lambda_\tau, \Sigma_{\omega,\tau+1}\}$ for $\tau < t$, but not on the history of the external state, or on the history of realizations of the memory noise $\{\omega_{t+1}\}$. In what follows, we therefore use the notation m_t for the vector of posterior means.

Among the state variables about which the memory state may convey information, we are particularly interested in the vector of variables $x_t = (\mu, y_{t-1})'$, which are the states determined prior to period t that are relevant for predicting the external state in periods $\tau \geq t$. Let $\bar{m}_t \equiv \mathrm{E}[x_t \,|m_t]$ be the two elements of the memory state that identify the posterior mean of x_t, and let Σ_t be the 2×2 block of second moments of x_t under this same posterior, so that

$$x_t \,|m_t \sim N(\bar{m}_t, \Sigma_t).$$

(Here $\bar{m}_t$ is now a 2-vector, and Σ_t a 2×2 matrix.) And let us furthermore introduce the vectors

$$e_1' \equiv [1 \quad 0], \qquad c' \equiv [1-\rho \quad \rho]$$

to select particular elements of this reduced state vector. Then $e_1'\bar{m}_t$ is the posterior mean and $e_1'\Sigma_t e_1$ the posterior variance for μ; while $c'\bar{m}_t$ is the posterior mean and $c'\Sigma_t c$ the posterior variance of the full-information forecast $E_{t-1}y_t$.

The for μ after also observing y_t will then be of the form (2.3), with mean and variance

given by the usual Kalman filter formulas,[18]

$$\hat{\mu}_t \;\equiv\; \mathrm{E}[\mu\,|s_t] \;=\; e_1'\bar{m}_t \,+\, \gamma_{1t}\,[y_t \,-\, c'\bar{m}_t], \tag{3.19}$$

$$\hat{\sigma}_t^2 \;\equiv\; \mathrm{var}[\mu\,|s_t] \;=\; e_1'\Sigma_t e_1 - \gamma_{1t}^2[c'\Sigma_t c \,+\, \sigma_\epsilon^2], \tag{3.20}$$

with a Kalman gain equal to[19]

$$\gamma_{1t} \;=\; \frac{e_1'\Sigma_t c}{c'\Sigma_t c \,+\, \sigma_\epsilon^2}. \tag{3.21}$$

Since y_t is observed precisely, these formulas completely characterize posterior beliefs in cognitive state s_t about the states x_{t+1} that are relevant for forecasting y_τ for all $\tau > t$. Note that $\hat{\sigma}_t^2$ is necessarily positive (complete certainty about the value of μ cannot be achieved in finite time, even in the case of perfect memory), and must satisfy the upper bound

$$\hat{\sigma}_t^2 \;\leq\; \hat{\sigma}_0^2 \;\equiv\; \frac{\Omega\sigma_y^2}{\Omega + \sigma_y^2}, \tag{3.22}$$

which corresponds to the degree of uncertainty about μ after observing the external state in the case of no informative memory whatsoever (the DM's situation in period $t = 0$).

Then the average losses from inaccurate forecasting in period t are given by

$$MSE_t \;=\; \hat{\sigma}_t^2. \tag{3.23}$$

This determines the value of one of the terms in (3.18) as a function of the posterior uncertainty associated with the memory state each period. We note that the optimal estimate $\hat{\mu}_t$ depends only on $\bar{m}_t$ (not other components of the memory state), and that the average loss implied by this estimate depends only on the posterior uncertainty Σ_t associated with those same two components.

[18] Note that these equations generalize (2.6)–(2.8) above for the $\rho = 0$ case.

[19] We use a 1 subscript in the notation for this variable because it is the first element of a vector of Kalman gains, defined in the more general formula given in Appendix B.

3.2 The sufficiency of memory of a reduced cognitive state

We further show in the appendix[20] that an optimal memory structure makes the memory state m_{t+1} a function only of the "reduced cognitive state"

$$\bar{s}_t \;\equiv\; \begin{bmatrix} \hat{\mu}_t \\ \\ y_t \end{bmatrix} \;=\; \mathrm{E}[x_{t+1}\,|s_t]. \tag{3.24}$$

We first note (using (3.19) and the fact that y_t is part of the cognitive state) that the elements of $\bar{s}_t$ are a linear function of s_t. Thus we can choose a representation of the vector s_t in which its elements are made up of two parts, $\bar{s}_t$ and $\underline{s}_t$, where the elements of $\underline{s}_t$ are uncorrelated with those of $\bar{s}_t$. We then observe that

$$\bar{m}_{t+1} \;=\; \mathrm{E}[\bar{s}_t\,|m_{t+1}].$$

Hence the only aspect of the memory state that matters for $\bar{m}_{t+1}$, and hence for determining both the optimal estimate $\hat{\mu}_{t+1}$ and the reduced cognitive state $\bar{s}_{t+1}$, will be the information that m_{t+1} contains about $\bar{s}_t$.

To the extent that m_{t+1} conveys any information about the elements of $\underline{s}_t$, this information has no consequences for the DM's estimates $\hat{\mu}_\tau$ in any periods $\tau \geq t+1$, but it increases the mutual information between s_t and m_{t+1}, and hence the information cost $c(I_t)$. Hence under an optimal information structure, the reduced memory state $\bar{m}_t$ must evolve according to a law of motion of the form

$$\bar{m}_{t+1} \;=\; \bar{\Lambda}_t \bar{s}_t \,+\, \bar{\omega}_{t+1}, \tag{3.25}$$

where $\bar{\omega}_{t+1} \;\sim\; N(0,\, \Sigma_{\bar{\omega},t+1})$ is distributed independently of the cognitive state. And in

[20] See Appendix C for details of the argument.

addition, the complete memory state must convey no more information about s_t than what is conveyed by the reduced memory state, so that we can without loss of generality assume that m_{t+1} consists solely of $\bar{m}_{t+1}$ (so that $d_{t+1} = 2$ for all $t \geq 0$).

Finally, the 2×2 matrices $\bar{\Lambda}_t$ and $\Sigma_{\bar{\omega},t+1}$ must satisfy additional restrictions in order for the reduced memory state defined in (3.25) to be consistent with the normalization

$$\mathrm{E}[\bar{s}_t \,|\bar{m}_{t+1}] \;=\; \bar{m}_{t+1}. \tag{3.26}$$

We show in the appendix that this relationship will hold if and only if[21]

$$\Sigma_{\bar{\omega},t+1} \;=\; (I - \bar{\Lambda}_t) X_t \bar{\Lambda}_t', \tag{3.27}$$

where $X_t \equiv \mathrm{var}[\bar{s}_t]$. Note that (3.24) implies that

$$\mathrm{var}[x_{t+1}] \;=\; \mathrm{var}[\bar{s}_t] \;+\; \mathrm{var}[x_{t+1} \,|s_t],$$

from which we see that

$$X_t \;=\; X(\hat{\sigma}_t^2) \;\equiv\; \begin{bmatrix} \Omega - \hat{\sigma}_t^2 & \Omega \\ \Omega & \Omega + \sigma_y^2 \end{bmatrix}. \tag{3.28}$$

Thus the matrix X_t depends only on the value of $\hat{\sigma}_t^2$. In addition, (3.22) implies that X_t will be positive semi-definite (p.s.d.), and non-singular (hence positive definite) except in the case that $\hat{\sigma}_t^2 = \hat{\sigma}_0^2$ (the case of a totally uninformative memory state m_t).

In order for it to be possible for (3.27) to hold, the matrix $\bar{\Lambda}_t$ must satisfy certain properties: (a) the matrix $\bar{\Lambda}_t X_t = X_t \bar{\Lambda}_t'$ must be symmetric (so that the right-hand side of (3.27) is also symmetric); and (b) the right-hand side of (3.27) must be a p.s.d. matrix.

[21] See the introductory section of Appendix D for details of the argument.

For any symmetric, positive definite 2×2 matrix X_t, we let $\mathcal{L}(X_t)$ be the set of matrices $\bar{\Lambda}_t$ with these properties. Then in addition to assuming that $\bar{\Lambda}_t \in \mathcal{L}(X_t)$, the variance matrix $\Sigma_{\bar{\omega},t+1}$ must be given by (3.27).

In the special case in which m_t is completely uninformative, $\hat{\mu}_t$ is proportional to the observation y_t, so that there exists a vector $w >> 0$ such that $\bar{s}_t = w \cdot y_t$. In this case,

$$X_t \;=\; X_0 \;\equiv\; [\Omega + \sigma_y^2]\, ww',$$

and we can show that the requirements stated above are satisfied by a matrix $\bar{\Lambda}_t$ if and only if $\bar{\Lambda}_t w = \lambda_t w$ (w is a right eigenvector), with an eigenvalue satisfying $0 \leq \lambda_t \leq 1$. Since the two elements of $\bar{s}_t$ are perfectly collinear in this case, the only part of the matrix $\bar{\Lambda}_t$ that matters for the evolution of the memory state is the implied vector $\bar{\Lambda}_t w$ (which must be a multiple of w). Thus we can without loss of generality impose the further restriction that if $\hat{\sigma}_t^2 = \hat{\sigma}_0^2$, we will describe the dynamics of the memory state using a matrix $\bar{\Lambda}_t$ of the form

$$\bar{\Lambda}_t \;=\; \lambda_t \frac{ww'}{w'w}, \tag{3.29}$$

for some $0 \leq \lambda_t \leq 1$. We now adopt this more restrictive definition of the set $\mathcal{L}(X_0)$ in this special case.[22]

We have now shown that the memory structure for period t is completely determined by a specification of a matrix $\bar{\Lambda}_t \in \mathcal{L}(X_t)$. In any period t, the value of $\hat{\sigma}_t^2$ and hence the matrix X_t will be implied by the choice of memory structure for the periods prior to t. Given a choice of $\bar{\Lambda}_t$, the variance-covariance matrix $\Sigma_{\bar{\omega},t+1}$ is uniquely determined by (3.27). As shown in the appendix,[23] this then uniquely determines Σ_{t+1}, indicating the degree of uncertainty implied by the memory state m_{t+1}, which then determines $\hat{\sigma}_{t+1}^2$ using (3.20).

[22]Restricting the set of transition matrices $\bar{\Lambda}_t$ that may be chosen in this way has no consequences for the evolution of the memory state, but it makes equation (3.30) below also valid in the case that $X_t = X_0$, and thus it allows us to state certain conditions below more compactly.

[23]See Appendix D.1 for details of the argument.

The degree of uncertainty about μ in the following period is then given by a function of the form

$$\hat{\sigma}_{t+1}^2 \;=\; f(\hat{\sigma}_t^2,\, \bar{\Lambda}_t),$$

that is uniquely defined for any non-negative $\hat{\sigma}_t^2$ satisfying the bound (3.22) and any $\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))$.

Then given that we start from an initial degree of uncertainty $\hat{\sigma}_0^2$ at time $t = 0$ defined by (3.22), we can define the class of sequences $\{\bar{\Lambda}_t\}$ for all $t \geq 0$ with the property that $\bar{\Lambda}_t \in \mathcal{L}(X_t)$ for all $t \geq 0$; let us call this class $\mathcal{L}^{seq}$. Moreover, for any sequence of transition matrices in $\mathcal{L}^{seq}$, we can uniquely define the sequences of values $\{\Sigma_t, \gamma_{1t}, \hat{\sigma}_t^2, X_t\}$ for all $t \geq 0$ implied by it. Thus given any sequence $\{\bar{\Lambda}_t\} \in \mathcal{L}^{seq}$, we can calculate the implied sequence of losses $\{MSE_t\}$ from forecast inaccuracy, using (3.23).

We can also uniquely identify the information cost implied by such a sequence of transition matrices, since as shown in the appendix,[24] the mutual information between s_t and m_{t+1} will be given by

$$I_t \;=\; I(\bar{\Lambda}_t) \;\equiv\; -\frac{1}{2}\log\det(I - \bar{\Lambda}_t) \tag{3.30}$$

each period. Note that the requirement that $\bar{\Lambda}_t \in \mathcal{L}(X_t)$ implies that

$$0 \;<\; \det(I - \bar{\Lambda}_t) \;\leq\; 1,$$

so that the quantity (3.30) is well-defined, and necessarily non-negative. As the elements of $\bar{\Lambda}_t$ are made small, so that memory ceases to be very informative about the prior cognitive state, $I - \bar{\Lambda}_t$ approaches the identity matrix, and I_t approaches zero. If $\bar{\Lambda}_t$ is varied in such a way as to make one of its eigenvalues approach 1, $I - \hat{\Lambda}_t$ approaches a singular matrix, and $\Sigma_{\hat{\omega},t+1}$ must approach a singular matrix as well; this means that in the limit, some linear combination of the elements of $\bar{s}_t$ is a random variable with positive variance that comes to be recalled with perfect precision. In this case, $\det(I - \hat{\Lambda}_t)$ approaches zero,

[24] See Appendix D.2 for details of the argument.

so that I_t grows without bound.

Thus a given sequence of transition matrices $\{\bar{\Lambda}_t\}$ uniquely determines sequences $\{MSE_t, I_t\}$, allowing the value of the objective (3.18) to be calculated. The problem of optimal design of a memory structure can then be reduced to the choice of a sequence $\{\bar{\Lambda}_t\} \in \mathcal{L}^{seq}$ so as to minimize (3.18). This objective is necessarily well-defined for any such sequence, since each of the terms is non-negative; the infinite sum will either converge to a finite value, or will diverge, in which case the sequence in question cannot be optimal.[25]

3.3 A recursive formulation

We now observe that if for any point in time t, we know the value of $\hat{\sigma}_t^2$ (which depends on the choices made regarding memory structure in periods $\tau < t$), the set of admissible transition matrices $\{\bar{\Lambda}_\tau\}$ for $\tau \geq t$ specifying the memory structure from that time onward will depend only on the value of $\hat{\sigma}_t^2$, and not any other aspect of choices made about the earlier periods. Moreover, any admissible continuation sequence $\{\bar{\Lambda}_\tau\}$ for $\tau \geq t$ implies unique continuation sequences $\{MSE_\tau, I_\tau\}$ for $\tau \geq t$, so that the value of the continuation objective

$$\sum_{\tau=t}^{\infty} \beta^{\tau-t} \left[\alpha \cdot MSE_\tau + c(I_\tau)\right] \tag{3.31}$$

will be well-defined.[26]

We can then consider the problem of choosing an admissible continuation plan $\{\bar{\Lambda}_\tau\}$ for $\tau \geq t$ so as to minimize (3.31), given an initial condition for $\hat{\sigma}_t^2$. (This is simply a more general form of our original problem choosing memory structures for all $t \geq 0$ to minimize (3.18), given the initial condition for $\hat{\sigma}_0^2$ specified in (3.22).) Let $V(\hat{\sigma}_t^2)$ be the lowest achievable value for (3.31), as a function of the initial condition $\hat{\sigma}_t^2$; this function is defined for any value of $\hat{\sigma}_t^2$ satisfying the bound (3.22), and is independent of the date t

[25]Note that it is clearly possible to choose memory structures for which the infinite sum converges. For example, if one chooses $\bar{\Lambda}_t = 0$ for all $t \geq 0$ (perfectly uninformative memory at all times), $MSE_t = \hat{\sigma}_0^2$ and $I_t = 0$ each period, and (3.18) will equal the finite quantity $\hat{\sigma}_0^2/(1-\beta)$.

[26]Since a finite value for the continuation objective is always possible (see (3.32) below), it is clear that plans that make (3.31) a divergent series cannot be optimal, and can be excluded from consideration.

from which we consider the continuation problem. Note that the lower bound necessarily lies in the interval

$$\alpha\hat{\sigma}_t^2 \;\leq\; V(\hat{\sigma}_t^2) \;\leq\; \alpha\left[\hat{\sigma}_t^2 \;+\; \frac{\beta}{1-\beta}\hat{\sigma}_0^2\right]. \tag{3.32}$$

(The lower bound follows from the fact that $MSE_t = \hat{\sigma}_t^2$, and all other terms in (3.31) must be non-negative; the upper bound is the value of (3.31) if one chooses $\bar{\Lambda}_\tau = 0$ for all $\tau \geq t$, which is among the admissible continuation plans.)

This value function also necessarily satisfies a Bellman equation of the form

$$V(\hat{\sigma}_t^2) \;=\; \min_{\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))} [\alpha\hat{\sigma}_t^2 \;+\; c(I(\bar{\Lambda}_t)) \;+\; \beta V(f(\hat{\sigma}_t^2, \bar{\Lambda}_t))], \tag{3.33}$$

where $I(\bar{\Lambda}_t)$ is the function defined in (3.30). Thus once we know how to compute the value function for arbitrary values of $\hat{\sigma}_{t+1}^2$, the problem of the optimal choice of a memory structure in any period t can be reduced to the one-period optimization problem stated on the right-hand side of (3.33). This indicates how the memory structure for period t must be chosen to trade off the cost $c(I_t)$ of retaining a more precise memory against the continuation loss $V(\hat{\sigma}_{t+1}^2)$ from having access to a less precise memory in period $t+1$.

Let $\mathcal{F}$ be the class of continuous functions $V(\hat{\sigma}_t^2)$, defined for values of $\hat{\sigma}_t^2$ consistent with (3.22), and consistent with the bounds (3.32) everywhere on this domain. Then (3.33) defines a mapping $\Phi : \mathcal{F} \to \mathcal{F}$: given any conjectured function $V(\hat{\sigma}_{t+1}^2) \in \mathcal{F}$ that is used to evaluate the right-hand side for any value of $\hat{\sigma}_t^2$, the minimized value of the problem on the right-hand side defines a new function $\tilde{V}(\hat{\sigma}_t^2)$ that must also belong to $\mathcal{F}$. Condition (3.33) states that the value function that defines the minimum achievable continuation loss must be a fixed point of this mapping: a function such that $V = \Phi(V)$.

We can further show that for any function $V \in \mathcal{F}$, the function $\Phi(V)$ defined by the right-hand side of (3.33) is necessarily a monotonically increasing function.[27] It follows that the fixed point $V(\hat{\sigma}_t^2)$ must be a monotonically increasing function. Moreover, we can

[27] See Appendix E.1 for a proof.

restrict the domain of the mapping Φ to the subset $\mathcal{F}^*$ of increasing functions.

This then provides us with an approach to computing the optimal memory structure for a given parameterization of our model. First, we find the value function $V(\hat{\sigma}^2) \in \mathcal{F}^*$ that is a fixed point of the mapping Φ, by iterating Φ to convergence. Then, given the value function, we can numerically solve the minimization problem on the right-hand side of (3.33) to determine the optimal transition matrix $\bar{\Lambda}_t$ in any period, once we know the value of $\hat{\sigma}_t^2$ for that period. Solution of this problem also allows us to determine the value of $\hat{\sigma}_{t+1}^2 = f(\hat{\sigma}_t^2, \bar{\Lambda}_t)$, so that the entire sequence of values $\{\hat{\sigma}_\tau^2\}$ for all $\tau \geq t$ can be determined once we know $\hat{\sigma}_t^2$. Finally, we recall that for the initial period $t = 0$, the value of $\hat{\sigma}_0^2$ is given by (3.22); we can thus solve for the entire sequence $\{\hat{\sigma}^2\}$ for all $t \geq 0$ by integrating forward from this initial condition.

Once we have determined the sequence of values $\{\hat{\sigma}_t^2\}$ implied by an optimal memory structure for each period, we can determine the elements of the matrix $X_t = X(\hat{\sigma}_t^2)$ each period, using (3.28). We show in the appendix[28] that the degree of uncertainty at the beginning of any period given the structure of the memory chosen for the previous period is given by

$$\Sigma_{t+1} \;=\; \Sigma_0 \;-\; X_t \bar{\Lambda}_t'.$$

This in turn allows us to calculate the DM's optimal estimate $\hat{\mu}_t$ each period, as a function of the history of realizations $\{y_\tau\}$ of the external state for all $0 \leq \tau \leq t$ and the history of realizations of the DM's memory noise $\{\tilde{\omega}_{\tau+1}\}$ for all $0 \leq \tau \leq t-1$, using (3.19). The DM's complete vector of forecasts z_t each period is then given by (3.16).

3.4 Optimality of a unidimensional memory state

We can show further that the optimal memory state must have a one-dimensional representation. This simplifies the computational formulation of the optimization problem on the right-hand side of (3.33), and provides further insight into the nature of an optimally

[28] See Appendix D.1 for details of the argument.

imprecise memory. Although the information contained in the cognitive state s_t that is relevant for predictiing (at time t) what actions will be desirable for the DM in later periods is two-dimensional (both elements of $\bar{s}_t$ matter, if $\rho > 0$, and except when memory is completely uninformative, these are not perfectly correlated with each other), we find that it is optimal for the DM's memory to include only a noisy record of a single linear combination of the two variables. Moreover, this is true regardless of how small memory costs may be.

There is in fact a fairly simple intuition for the result. Note that in any period t, the Kalman filter (3.19) implies that the optimal estimate of the unknown value of μ will be given by a linear function of elements of the cognitive state of the form

$$\hat{\mu}_t \;=\; \zeta_t \;+\; \delta_t' \bar{m}_t. \tag{3.34}$$

It follows from this that the only information in the memory state m_t that matters for the estimate $\hat{\mu}_t$ is the single quantity $\delta_t' \bar{m}_t$.

We can establish the optimality of a unidimensional memory in the following way. Consider the optimization problem on the right-hand side of (3.33) in any period t, given the degree of uncertainty $\hat{\sigma}_t^2$ determined by the memory structures chosen in earlier periods. The fact that $V(\hat{\sigma}_{t+1}^2)$ is an increasing function, and that $c(I_t)$ is at least weakly increasing, means that an optimal memory structure must minimize the mutual information I_t given the uncertainty $\hat{\sigma}_{t+1}^2$ that it implies for the following period.[29] Hence the optimal choice for $\bar{\Lambda}_t$ must solve the problem

$$\min_{\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))} I(\bar{\Lambda}_t) \qquad \text{s.t. } \; f(\hat{\sigma}_t^2, \bar{\Lambda}_t) \;\leq\; \hat{\sigma}_{t+1}^2, \tag{3.35}$$

[29] In the case that $c(I)$ is constant over some interval, reducing I_t need not reduce $c(I_t)$, but it cannot increase it; thus the solution to the problem (3.35) must be among the solutions to the problem on the right-hand side of (3.33), even if it is not a unique solution. In such case, showing that the solution to (3.35) is necessarily a singular matrix suffices to show that we can without any loss impose the further constraint in (3.33) that the matrix $\bar{\Lambda}_t$ must be at most of rank one.

for given values of $(\hat{\sigma}_t^2, \hat{\sigma}_{t+1}^2)$. We shall show that whenever $(\hat{\sigma}_t^2, \hat{\sigma}_{t+1}^2)$ are such that the set of matrices satisfying the constraint in (3.35) is non-empty,[30] the solution $\bar{\Lambda}_t$ to this problem must be at most of rank one. Thus it must be of the special form

$$\bar{\Lambda}_t \;=\; \lambda_t X_t v_t v_t', \tag{3.36}$$

where λ_t is a scalar satisfying $0 \leq \lambda \leq 1$ and v_t is a vector normalized to satisfy $v_t' X_t v_t = 1$. It follows that in each period $\bar{m}_{t+1} \;=\; X_t v_t \tilde{m}_{t+1}$, where $\tilde{m}_{t+1}$ is a unidimensional memory state with a law of motion

$$\tilde{m}_{t+1} \;=\; \lambda_t v_t' \bar{s}_t \;+\; \tilde{\omega}_{t+1}, \qquad \tilde{\omega}_{t+1} \;\sim\; N(0,\, \lambda_t(1-\lambda_t)). \tag{3.37}$$

If $\hat{\sigma}_t^2 = \hat{\sigma}_0^2$, the set $\mathcal{L}(X_0)$ consists only of matrices of the form (3.36), with

$$v_t \;=\; \frac{w}{(\Omega + \sigma_y^2)^{1/2}(w'w)}, \tag{3.38}$$

because of (3.29). Hence the asserted result is obviously true in that case. Suppose instead that $\hat{\sigma}_t^2 < \hat{\sigma}_0^2$, and consider any matrix $\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))$ that satisfies the constraint in (3.35). If $\bar{\Lambda}_t$ is not itself of rank one (or lower), we shall show that we can choose an alternative transition matrix of the form (3.36), that is also consistent with the constraint in (3.35), but which achieves a lower value of $I(\bar{\Lambda}_t)$.

Let the alternative transition matrix be given by (3.36), with

$$\lambda_t \;=\; \frac{\delta_{t+1}' \bar{\Lambda}_t X_t \bar{\Lambda}_t' \delta_{t+1}}{\delta_{t+1}' X_t \bar{\Lambda}_t' \delta_{t+1}}, \qquad v_t \;=\; \frac{\bar{\Lambda}_t' \delta_{t+1}}{(\delta_{t+1}' \bar{\Lambda}_t X_t \bar{\Lambda}_t' \delta_{t+1})^{1/2}},$$

where $\delta_{t+1} \equiv e_1 \;-\; \gamma_{1,t+1} c$ is the vector introduced in (3.34), and let the matrix $\Sigma_{\bar{\omega},t+1}$ be correspondingly modified in the way specified by (3.27). We show in the appendix[31] that

[30] Note that this must be the case if $\hat{\sigma}_{t+1}^2$ is chosen optimally given $\hat{\sigma}_t^2$.

[31] See Appendix E.2 for details of the argument.

this specification implies that $0 \leq \lambda_t \leq 1$, so that this alternative matrix also belongs to $\mathcal{L}(X_t)$. Moreover, the new memory structure implies a conditional distribution

$$\delta'_{t+1}\bar{m}_{t+1}|s_t \ \sim \ N(\delta'_{t+1}\bar{\Lambda}_t\bar{s}_t,\ \delta'_{t+1}\Sigma_{\bar{\omega},t+1}\delta_{t+1})$$

that is the same as under the original memory structure. This implies that the optimal estimate $\hat{\mu}_{t+1}$ conditional on the cognitive state s_{t+1} will be the same function of $\bar{m}_{t+1}$ and y_{t+1} in the case of the new memory structure, and that the conditional distribution $\hat{\mu}_{t+1}|s_t, y_{t+1}$ will be the same. It follows that $\hat{\sigma}^2_{t+1}$ will be the same, so that the alternative transition matrix also satisfies the constraint in (3.35).

At the same time, we show in the appendix that the reduction in the complexity of memory cannot increase information costs in any period.[32] The new memory structure consists effectively of a scalar memory state $\tilde{m}_{t+1}$ in each period, which is a multiple of $d'_{t+1}\bar{m}_{t+1}$, a particular linear combination of the elements of the memory state under the previous memory structure. Hence the information about $\bar{s}_t$ that is revealed by m_{t+1} under the new memory structure (i.e., that is revealed by $\tilde{m}_{t+1}$) is also information that was revealed by $\bar{m}_{t+1}$ under the previous memory structure; thus the value of I_t under the previous memory structure must have been at least as large as under the new memory structure. In fact, the only case in which the mutual information will not be reduced by the proposed modification of the memory structure is if all elements of $\bar{m}_{t+1}$ were multiples of $d'_{t+1}\bar{m}_{t+1}$; which is to say, only if $\bar{\Lambda}_t$ were already of the special form (3.36).

We conclude, then, that an optimal memory structure must involve a transition matrix in every period of the special form (3.36), so that the memory state each period can be represented by a scalar quantity $\tilde{m}_t$. The choice of memory structure can then be reduced to a problem of choosing, in each period $t \geq 0$, a scalar quantity $0 \leq \lambda_t \leq 1$, and the direction of a vector v_t (the length of which will then be chosen each period so as to ensure that $v'_t X_t v_t = 1$); the values chosen for these quantities then determine the law of motion for

[32] See Appendix E.2 for details of the argument.

the unidimensional memory state $\tilde{m}_{t+1}$, specified by (3.37). This in turn determines the elements of the matrix Σ_{t+1}, and hence the value of the gain coefficient $\gamma_{1,t+1}$ in the Kalman filter formula (3.19) and the value of $\hat{\sigma}^2_{t+1}$, which determines the matrix $X_{t+1} = X(\hat{\sigma}^2_{t+1})$.

For any value $0 \leq \hat{\sigma}^2_t < \hat{\sigma}^2_0$, let $\mathcal{V}(\hat{\sigma}^2_t)$ be the set of vectors v_t satisfying $v_t' X(\hat{\sigma}^2_t) v_t = 1$. In the case that $\hat{\sigma}^2_t = \hat{\sigma}^2_0$, we add the further stipulation that $\mathcal{V}(\hat{\sigma}^2_0)$ consists only of the single vector (3.38). Then given a value for $\hat{\sigma}^2_t$, determined by the memory structures for periods $\tau < t$, the memory structure for period t is specified by a scalar quantity $0 \leq \lambda_t \leq 1$ and a vector $v_t \in \mathcal{V}(\hat{\sigma}^2_t)$. These determine a value for $\hat{\sigma}^2_{t+1} = f(\hat{\sigma}^2_t, \lambda_t, v_t)$, where now the function f is defined for any values of its arguments satisfying $0 \leq \hat{\sigma}^2_t \leq \hat{\sigma}^2_0$, $0 \leq \lambda_t \leq 1$, and $v_t \in \mathcal{V}(\hat{\sigma}^2_t)$.

Because of the monotonicity of the value function $V(\hat{\sigma}^2_{t+1})$, the optimal weight vector v_t in any period must be the one that solves the static optimization problem

$$\bar{f}(\hat{\sigma}^2_t, \lambda_t) \;\equiv\; \min_{v_t \in \mathcal{V}(\hat{\sigma}^2_t)} f(\hat{\sigma}^2_t, \lambda_t, v_t). \tag{3.39}$$

In the appendix,[33] we give an explicit algebraic solution for the optimal v_t for any given values $0 \leq \hat{\sigma}^2_t \leq \sigma^2_0$ and $0 < \lambda_t \leq 1$,[34] and hence for the function $\bar{f}(\hat{\sigma}^2_t, \lambda_t)$. The latter function is also defined when $\lambda_t = 0$, and easily seen to equal $\bar{f}(\hat{\sigma}^2_t, 0) = \hat{\sigma}^2_0$. Thus we can solve for the dynamics of $\{\hat{\sigma}^2_t\}$ implied by any sequence $\{\lambda_t\}$, by iterating the law of motion

$$\hat{\sigma}^2_{t+1} \;=\; \bar{f}(\hat{\sigma}^2_t, \lambda_t),$$

starting from the initial condition $\hat{\sigma}^2_0$ defined in (3.22).

Moreover, it follows from (3.30) that the mutual information associated with the period t memory structure will be given by

$$I_t \;=\; -\frac{1}{2}\log(1 - \lambda_t), \tag{3.40}$$

[33] See Appendix E.3 for details.

[34] Note that no solution is needed in the case that $\lambda_t = 0$, since in this case v_t is undefined.

just as in the i.i.d. case discussed in section 1. The Bellman equation (3.33) can therefore be written in the simpler form

$$V(\hat{\sigma}_t^2) \;=\; \min_{0\leq\lambda_t\leq 1}[\alpha\hat{\sigma}_t^2 \;+\; c(-(1/2)\log(1-\lambda_t)) \;+\; \beta V(\bar{f}(\hat{\sigma}_t^2,\lambda_t))]. \tag{3.41}$$

4 Features of the Model Solution

Here we provide numerical examples of solutions for an optimal memory structure, under alternative assumptions about both the degree of persistence of the process that must be forecasted and the nature of the information cost function. In reporting our results, it is useful to describe the model solution in terms of scale-invariant quantities — that is, ones that are independent of the value of σ_y, indicating the amplitude of the transitory fluctuations in the external state around its mean. Thus we parameterize the degree of prior uncertainty about μ not in terms a value for Ω (the variance of the prior distribution for μ), but rather by a value for $K \equiv \Omega/\sigma_y^2$ (the variance of the prior distribution for μ/σ_y). We similarly measure the degree of uncertainty about μ conditional on the cognitive state at a given point in time (i.e., after a given amount of experience) not in terms of the value of $\hat{\sigma}_t^2$, but rather by the scaled uncertainty measure $\eta_t \equiv \hat{\sigma}_t^2/\sigma_y^2$.

In terms of this scaled uncertainty measure, an optimal memory structure minimizes the value of

$$\sum_{t=0}^{\infty} \beta^t \,[\eta_t \;+\; \tilde{c}(I_t),]$$

a scaled version of (3.18), where the scaled cost function is defined as $\tilde{c}(I) \equiv c(I)/(\alpha\sigma_y^2)$. (Dividing by α further reduces the numbers of parameters that we need to specify in considering the different possible forms that the optimal memory structure may take, since it is only the relative weights on the two loss terms in the objective (3.18) that matter for the optimal memory structure.) Our scale-invariant model is then completely specified by values for the parameters ρ, β, K and the scaled cost function $\tilde{c}(I)$. In our quantita-

tive analysis, we assume that each "period" of our discrete-time model corresponds to a quarter of a year (the variable to be forecasted is a quarterly time series), and hence set $\beta = 0.99$ (implying a discount rate of 4 percent per annum). We consider a variety of values $0 \leq \rho < 1$ for the assumed degree of serial correlation of the external state, and explore the effects of different assumptions regarding the degree of prior uncertainty and the size of information costs.

4.1 The case of a fixed per-period bound on mutual information

We begin by considering the case in which $\tilde{c}(I) = 0$ for all $I \leq \bar{I}$, but values of I_t greater than $\bar{I}$ are infeasible, as assumed in section 1. Solution for the optimal memory structure is particularly simple in this case. Because of (3.40), the per-period bound on mutual information can equivalently be written as an upper bound $\lambda_t \leq \bar{\lambda}$, just as in section 1. The optimal memory structure in period t is then characterized by the λ_t that minimizes $\bar{f}(\hat{\sigma}_t^2, \lambda_t)$ subject to this constraint. We show in the appendix[35] that the minimizing value of λ_t is necessarily the largest feasible value; hence in the solution to this problem, $\lambda_t = \bar{\lambda}$, the value determined by the per-period information bound.

The dynamics of the uncertainty measure are then given by $\hat{\sigma}_{t+1}^2 = \bar{f}(\hat{\sigma}_t^2, \bar{\lambda})$. In terms of the rescaled variables, the law of motion can be written as

$$\eta_{t+1} \;=\; \phi(\eta_t; \bar{\lambda}), \tag{4.42}$$

where $\phi(\eta; \bar{\lambda})$ is a function that is independent of the scale factor σ_y.[36]

For any value of $\bar{\lambda}$ indicating the tightness of the constraint on the complexity of memory, equation (4.42) indicates how the DM's degree of uncertainty about μ evolves as additional observations of the external state are made. Starting from the initial condition $\eta_0 = K/(K+1)$ implied by (3.22), the law of motion (4.42) can be iterated to obtain a

[35] See Appendix F.1 for details of the argument.

[36] See Appendix F.1 for an explicit algebraic solution for this function.

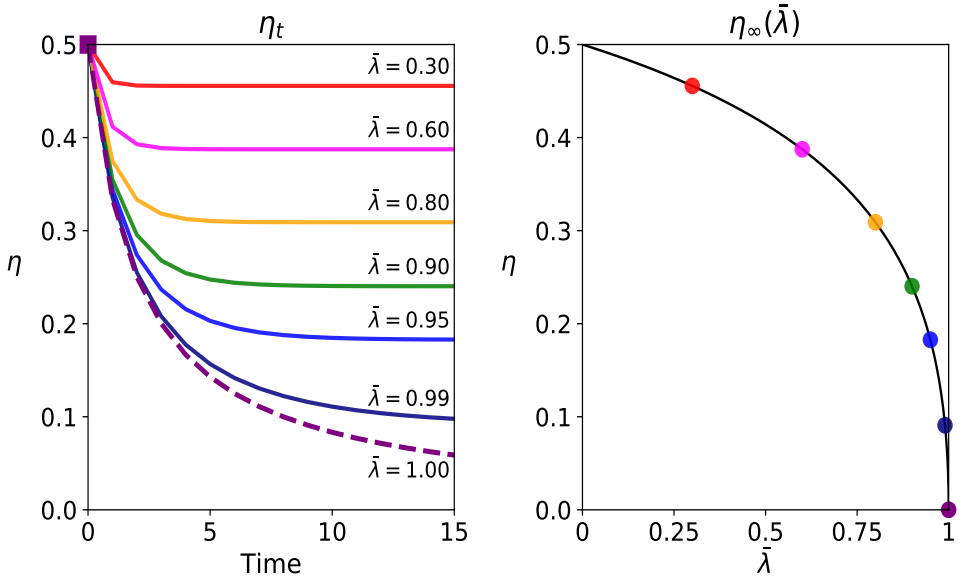


Figure 1.1: The evolution of uncertainty about μ (for varying degress of $\bar{\lambda}$)

The evolution of scaled uncertainty about μ as the number t of previous (imperfectly remembered) observations grows. The right panel shows the long-run value of scaled uncertainty (to which η_t converges as $t \to \infty$) as a function of the constraint on the complexity of memory, parameterized by $\bar{\lambda}$.

unique solution for the complete sequence of values $\{\eta_t\}$ for all $t \geq 0$. In the limiting case $\bar{\lambda} = 1$ (unlimited memory), the law of motion (4.42) takes the especially simple form

$$\frac{1}{\eta_{t+1}} = \frac{1}{\eta_t} + \frac{1-\rho}{1+\rho}. \tag{4.43}$$

This is simply the standard result for the linear growth in posterior precision under Bayesian updating as additional observations are made; it has the implication that η_t declines monotonically, and converges to zero for large t. Thus in the case of perfect memory, the DM should eventually learn the value of μ with perfect precision, and hence make forecasts of the kind implied by the hypothesis of rational expectations.

When $\bar{\lambda} > 0$, instead, the law of motion (4.42) implies that η_t should decrease initially, as even imprecise memory of the DM's observations of the external state reduces uncertainty to some degree, but that η_t remains bounded away from zero, and converges to a value $\eta_\infty(\bar{\lambda}) > 0$. This is illustrated in Figure 1, which shows the dynamics implied by

(4.42) for each of several different values of $\bar{\lambda}$, in the case that $\rho = 0$ and $K = 1$.[37] The left panel plots the sequence of values $\{\eta_t\}$ implied by (4.42) for a given value of $\bar{\lambda}$. (Note that the initial value η_0 is the same in each case.) The right panel shows the value η_∞ to which the sequence converges as t grows; this value depends on $\bar{\lambda}$, and the functional relationship between $\bar{\lambda}$ and this limiting degree of uncertainty can be described by a function $\eta_\infty(\bar{\lambda})$, plotted as a smooth curve in the right panel of the figure.

In the case that $\bar{\lambda} = 1$ (shown as a dashed curve in the left panel of Figure 1), the sequence $\{\eta_t\}$ decreases monotonically to zero at the rate predicted by the difference equation (4.43). But for any number of prior observations $t > 0$, the value of η_t remains higher the lower is $\bar{\lambda}$ (that is, the tighter the memory constraint), and the long-run degree of uncertainty about μ, measured by η_∞, is a decreasing function of $\bar{\lambda}$ as well, as shown by the curve in the right panel of the figure. Because of the limit on the amount of information that can be retained in memory, the DM's uncertainty about the value of μ never falls below a certain level, even in the long run, despite our assumption that the value of μ is fixed for all time. We further observe that the long-run degree of uncertainty η_∞ is larger, the smaller is $\bar{\lambda}$ (that is, the tighter the constraint on memory). In the limit as $\bar{\lambda}$ approaches zero (corresponding to a constraint that memory must be completely uninformative), the long-run degree of uncertainty η_∞ approaches the prior degree of uncertainty $\eta_0 = K/(K+1)$.

As η_t falls along one of these trajectories, the weight vector v_t that solves the problem (3.39) shifts as well. As η_t converges to the long-run value η_∞, the optimal weight vector v_t similarly converges to a long-run value v_∞, indicating the particular linear combination of $\hat{\mu}_t$ and y_t that is imprecisely recorded in memory each period. Associated with this stationary long-run memory structure there will also be a stationary long-run value for the Kalman gain coefficient γ_1 in equation (3.19), and more generally, stationary values for

[37] The effects of variation in the parameters ρ and K are illustrated in additional figures shown in Appendix F.1. We use the parameterization $K = 1$ in the figures shown in the text because this value allows a reasonably good fit of the predictions shown in Figure 7 below with the experimental evidence reported by Afrouzi *et al.* (2020).

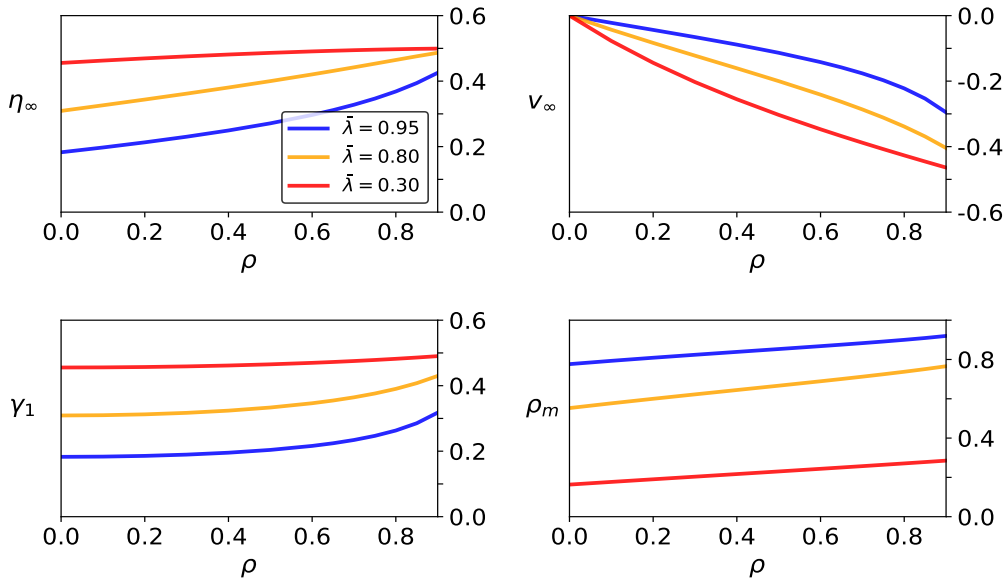


Figure 1.2: The optimal memory structure in the long run (for varying degress of $\bar{\lambda}$)

Coefficients describing the optimal memory structure in the long run, as a function of the degree of persistence ρ of the external state, for alternative values of $\bar{\lambda}$. Respective panels show the long-run values for η (measuring uncertainty about μ), the direction vector v (indicating the content of the memory state), the Kalman gain γ_1 (for updating the DM's estimate of μ), and ρ_m (measuring the intrinsic persistence of fluctuations in the memory state).

the coefficients of the linear difference equations describing the joint dynamics $\{y_t, \tilde{m}_t\}$ of the external state and the memory state.

These long-run stationary coefficients will depend on the value of $\bar{\lambda}$ (indicating the tightness of the memory constraint) and also on the value of ρ (indicating the degree of persistence of the fluctuations in the external state). Figure 2 indicates how variation in each of these parameters affects several of the long-run stationary coefficients.[38] In each panel, a curve shows how the coefficient in question varies as a function of ρ (for values of ρ between 0.0 and 0.9), for a given value of $\bar{\lambda}$; curves of this kind are shown for each of three different values of $\bar{\lambda}$, ranging between $\bar{\lambda} = 0.95$ (in which case memory is relatively precise) and $\bar{\lambda} = 0.30$ (in which case it is much more constrained). All of the curves shown in Figure 2 are again for the case of prior uncertainty $K = 1$.

[38] See Appendix G.1 for the formulas used to calculate each of the coefficients plotted here as functions of the model parameters.

The upper-right panel of the figure shows the long-run direction vector v_∞; the quantity reported on the vertical axis is the long-run value of the ratio v_2/v_1 of the vector's two components.[39] Thus a value of -0.3 for this quantity means that the univariate memory state $\tilde{m}_{t+1}$ is (up to a multiplicative factor that does not affect its information content) equal to the value of $\hat{\mu}_t - 0.3y_t$, plus additive Gaussian noise. The figure shows that when $\rho = 0$, the optimal univariate memory state involves $v_2 = 0$; that is, only the current estimate $\hat{\mu}_t$ of the unknown mean is remembered with noise, with the current observation y_t being completely forgotten. This is optimal because when $\rho = 0$, the current value y_t contains no information that is relevant for improving subsequent forecasts of the external state, except to the extent that it helps to improve the DM's estimate of μ (which information is already reflected in the estimate $\hat{\mu}_t$). Instead, when the external state is serially correlated, it is optimal to commit to memory a linear combination of $\hat{\mu}_t$ and the current state y_t; in the case that $\rho > 0$, the optimal linear combination puts a negative relative weight on y_t, to an extent that is greater the greater the degree of serial correlation, and greater the tighter the constraint on memory.

The upper-left panel of the figure shows the long-run degree of uncertainty about μ, measured by η_∞. As shown in Figure 1, when $\rho = 0$, η_∞ is a decreasing function of $\bar{\lambda}$. We see in Figure 2 that this is also true when $\rho > 0$. However, for a given memory constraint $\bar{\lambda}$, the long-run value η_∞ is also an increasing function of ρ, with the degree of increase when the external state is highly persistent being particularly notable when memory is more accurate. The greater the serial correlation of the state, the fewer the effective number of independent noisy observations of μ that the DM receives over any finite time period; thus even under perfect Bayesian updating, equation (4.43) indicates that the rate at which precision is increased by each additional observation is smaller the larger is ρ.

[39] This information (together with the value of η_∞ given in the upper left panel) suffices to completely determine the vector v_t, since the vector is normalized so that $v'Xv = 1$. The value of λ (given by the constraint $\bar{\lambda}$), the matrix X (determined by the value of η_∞), and the vector v then completely determine the long-run stationary elements of the matrix $\bar{\Lambda}$ (using (3.36)) and hence also of the matrix $\Sigma_{\bar{\omega}}$ (using (3.27)); thus the dynamics of the memory state given by (3.25) are completely specified.

In the case of perfect memory, the long-run degree of uncertainty about μ is nonetheless zero (there is simply slower convergence to that long-run value when ρ is large); but with moderately imperfect memory, the effective amount of experience that can ever be drawn upon remains bounded, so that the uncertainty about μ remains larger forever when ρ is larger. When memory is even more imperfect, not much more than one observation (the most recent one) can be used in any event, so that the value of η_∞ is in this case less sensitive to the value of ρ.

In the long run, the dynamics of the cognitive state $\bar{s}_t$ and the memory state $\bar{m}_{t+1}$ are described by linear equations with constant coefficients. The lower-left panel of Figure 2 shows the long-run value for the Kalman gain γ_{1t} in (3.19). With imperfect memory, this is always a quantity between 0 and 1, meaning that a higher value of the current state y_t raises the DM's estimate of the value of μ, though by less than the amount of the increase in y_t. For a given value of ρ, the Kalman gain is larger the tighter the constraint on memory; in the limit as $\bar{\lambda} \to 1$ (perfect memory), the long-run value of this coefficient approaches zero (as the true value of μ is eventually learned), while in the limit as $\bar{\lambda} \to 0$ (no memory), the value approaches a maximum value that is still less than one (because of the DM's finite-variance prior).

Finally, the lower-right panel reports the long-run value of ρ_m, a measure of the intrinsic persistence of the memory state. The impulse response function for the effect of a memory-noise innovation $\tilde{\omega}_t$ on the subsequent path of the univariate memory state $\tilde{m}_\tau$ is proportional to $(\rho_m)^{\tau-t}$ for all $\tau \geq t$;[40] thus the value of ρ_m indicates the rate of exponential decay of the memory state back to its long-run average value. A smaller value of ρ_m means that the contents of memory decay more rapidly; for any value of ρ, the figure shows that ρ_m is smaller, the tighter the memory constraint. At the same time, while a larger value of ρ_m implies that memory persists for a longer time, it also implies that when memory

[40]Here we refer to the difference that the realization of $\tilde{\omega}_t$ makes for the forecasts of $\tilde{m}_\tau$ at different horizons $\tau \geq t$, by an observer who knows the true value of μ and the DM's cognitive state at time $t-1$, in addition to observing the realization of $\tilde{\omega}_t$. See Appendix G.1 for details of the calculation.

noise creates an erroneous impression of prior experience, this bias in what is recalled about is also corrected more slowly; thus the value of ρ_m is an important determinant of the predicted persistence of forecast bias.

4.2 The case of a linear cost of information

Analysis of the model is more complex when instead the amount of information stored in memory each period can be increased at some finite cost. As an illustration we consider the polar opposite case in which $\tilde{c}(I)$ is a linear function of I, so that the marginal cost of a further increase in the mutual information is independent of how large it already is. Thus we assume that $\tilde{c}(I) = \tilde{\theta} \cdot I$, for some coefficient $\tilde{\theta} > 0$ which parameterizes the cost of memory.

In this case, the optimal choice of λ_t in any period will depend on the value of reducing uncertainty in the following period. We note that the value function $V(\hat{\sigma}^2_{t+1})$ appearing in the Bellman equation (3.41) can be written as $\sigma_y \cdot \tilde{V}(\eta_{t+1})$, where η_{t+1} is the scaled uncertainty measure and the function $\tilde{V}(\eta)$ is independent of the scale factor σ_y (for given values of the parameters K, ρ, β and $\tilde{\theta}$). We can then write the Bellman equation in the scale-invariant form

$$\tilde{V}(\eta_t) \;=\; \min_{0 \leq \lambda_t \leq 1} \left\{ \eta_t \;-\; \frac{\tilde{\theta}}{2} \log(1 - \lambda_t) \;+\; \beta \tilde{V}(\phi(\eta_t; \lambda_t)) \right\}. \tag{4.44}$$

The optimal choice of λ_t in any period will be the value that solves the problem on the right-hand side of (4.44). This problem has a solution $\lambda_t = \lambda^*(\eta_t)$ which depends only on the value of η_t, the degree of uncertainty in period t determined by the memory structures chosen for periods prior to t.

Thus we can solve for the optimal policy function $\lambda^*(\eta_t)$ once we know the value function $\tilde{V}(\eta_{t+1})$, and we can solve numerically for the value function by iterating the Bellman equation (4.44), as discussed further in the appendix.[41] The policy function $\lambda_t = \lambda^*(\eta_t)$

[41] See Appendix F.2 for details.

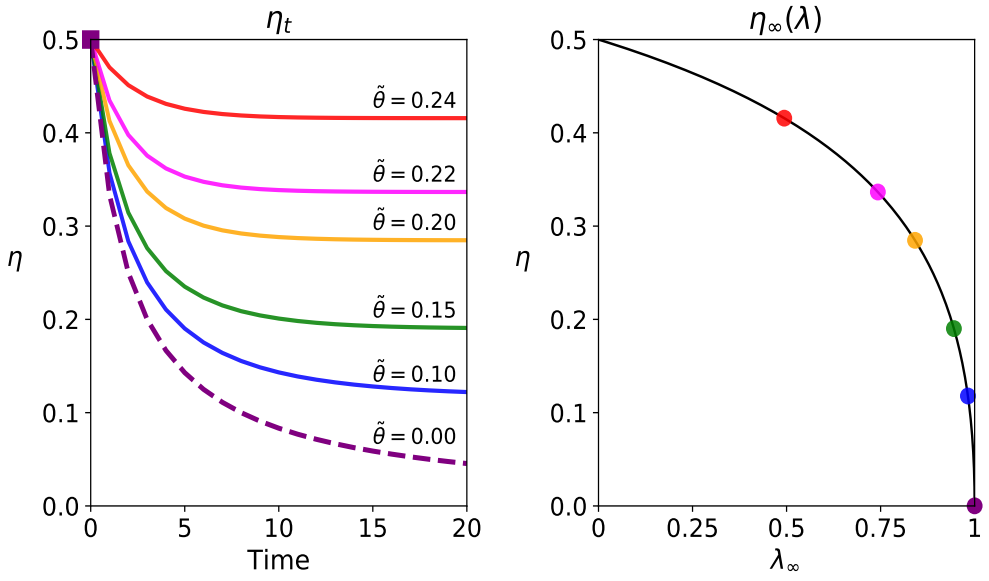


Figure 1.3: The evolution of uncertainty about μ (for varying degress of $\tilde{\theta}$)

The evolution of scaled uncertainty about μ as the number t of previous (imperfectly remembered) observations grows, now for the case of a linear cost of memory complexity. The right panel shows the long-run value of scaled uncertainty for each value of the cost parameter $\tilde{\theta}$, plotted as a point on the same locus of optimal long-run memory structures as in Figure 1.

together with the law of motion

$$\eta_{t+1} \;=\; \phi(\eta_t; \lambda_t) \tag{4.45}$$

derived earlier can then be solved for the dynamics of the scaled uncertainty $\{\eta_t\}$ for all $t \geq 0$, starting from the initial condition $\eta_0 = K/(K+1)$.[42] The dynamics of scaled uncertainty as a function of the number of observations t are shown for progressively larger values of $\tilde{\theta}$ in Figure 3, using the same format as in Figure 1. Once again, we see that while uncertainty about μ eventually falls to zero as a result of when there is no cost of memory complexity, as long as the cost is positive, the value of η_t remains bounded away from zero, and converges asymptotically to a value η_∞ that is higher the higher the cost of memory complexity.

Associated with such an asymptotic degree of uncertainty is a particular long-run memory structure $(\lambda_\infty, v_\infty)$, which will imply a particular long-run value for the Kalman gain

[42] See Appendix F.2 for further discussion of the implied dynamics.

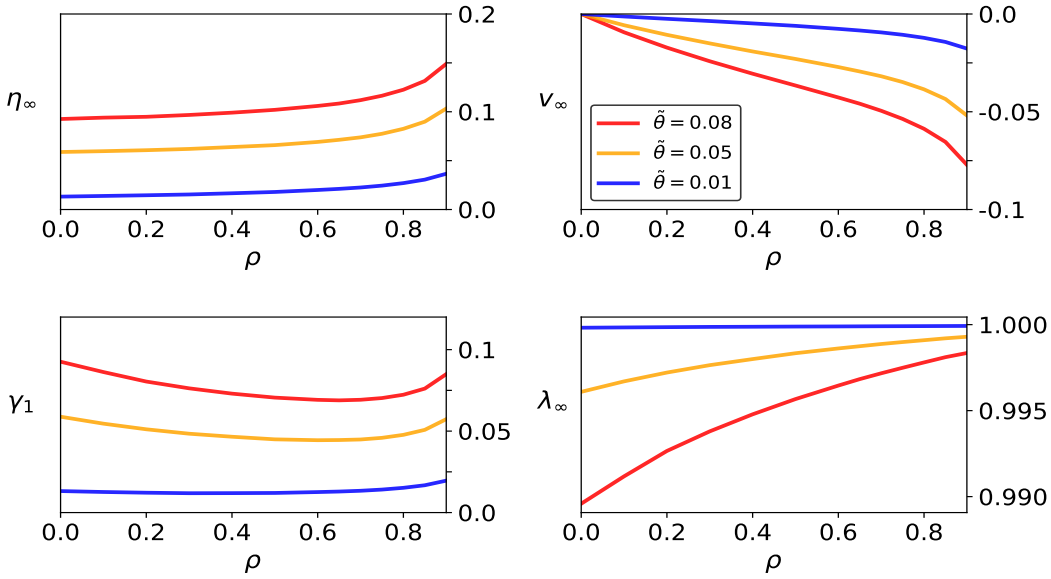


Figure 1.4: The optimal memory structure in the long run (for varying degress of $\tilde{\theta}$)

Coefficients describing the optimal memory structure in the long run, as a function of the degree of persistence ρ of the external state, in the case of a linear memory cost function, for alternative values of $\tilde{\theta}$. Respective panels show the long-run values for η, the direction vector v, the Kalman gain γ_1, and the memory precision coefficient λ.

γ_1. The way in which the long-run values of these different quantities vary with different assumptions about the values of ρ and $\tilde{\theta}$ is illustrated in Figure 4. (We use the same convention as in Figure 2 to indicate the direction of the vector v_∞ in the upper-right panel of the figure.) As we vary ρ for a given value of $\tilde{\theta}$, the associated value of λ_∞ changes; hence the fixed-$\tilde{\theta}$ curves shown in Figure 4 do not correspond exactly to any of the fixed-λ curves plotted in Figure 2, even though each of the long-run memory structures associated with a pair $(\rho, \tilde{\theta})$ is identical to the long-run memory structure associated with some pair $(\rho, \bar{\lambda})$. As shown in the lower-right panel of the figure, the optimal λ_∞ rises as ρ increases, for any value of the cost parameter $\tilde{\theta} > 0$; the more persistent the external state that must be forecasted, the more it becomes worthwhile to pay a larger information cost in order to retain a more precise memory of prior experience.

Not surprisingly, we observe that for any value of ρ, increasing the memory cost $\tilde{\theta}$ makes it optimal for the long-run precision of memory λ_∞ to be smaller, and consequently for the

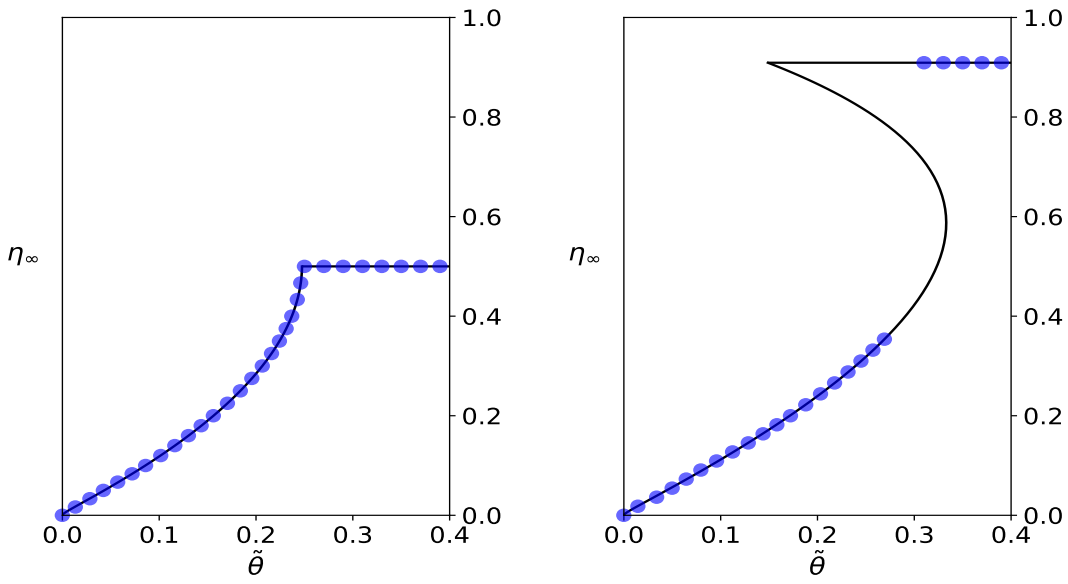


Figure 1.5: The long-run uncertainty

Long-run value of the scaled uncertainty measure η_∞ (blue dots) as a function of the cost parameter $\tilde{\theta}$, in the case of a linear memory cost function. Left panel: $K = 1, \rho = 0$. Right panel: $K = 10, \rho = 0$.

long-run degree of uncertainty about μ to be larger. In the case of a sufficiently high value of $\tilde{\theta}$, it will be optimal for memory to be completely uninformative. In fact, this happens for a finite value of $\tilde{\theta}$, and it occurs abruptly, rather than through a gradual increase in the long-run degree of uncertainty η_∞ toward the limiting value of $\eta_0 = K/(K+1)$ as $\tilde{\theta}$ is increased. A graph of the relationship between η_∞ and the value of $\tilde{\theta}$ is shown in Figure 5, for the case $\rho = 0$, and two different possible values of K: $K = 1$ and $K = 10$. For each value of $\tilde{\theta}$, the value of η_∞ associated with the optimal memory structure is shown by a large blue dot.

In each panel of this figure, the continuous black curve is the correspondence consisting of all points $(\tilde{\theta}, \eta_\infty)$ such that η_∞ is a stationary solution of the Euler equation associated with the optimization problem on the right-hand side of (4.44).[43] The Euler equation represents a first-order condition for the optimal choice of the degree of precision of memory; satisfaction of this condition is necessary but not sufficient for memory precision leading

[43] See Appendix F.4 for derivation of this equation.

to $\eta_{t+1} = \eta$ to be optimal starting from a situation in which $\eta_t = \eta$. Because the objective function on the right-hand side of (4.44) is not a convex function, it can have multiple local minima (as well as a local maximum located between two local minima). Which of the local minima represents the global minimum (and hence the optimal memory structure) can jump abruptly as a result of a small change in parameters;[44] this is what happens when the value of η_∞ changes abruptly in the right panel of Figure 5, for a value of $\tilde{\theta}$ slightly above 0.28.

In the $K = 10$ case, we see that there need not be a unique value of η_∞ for a given value of $\tilde{\theta}$ that represents a stationary solution to the Euler equation. For any value of $\tilde{\theta}$ greater than a critical value around 0.15, if one starts from $\eta_t = \eta_0$ (a completely uninformative memory), the choice of $\eta_{t+1} = \eta_0$ again represents a local minimum of the objective; hence $\eta = \eta_0$ is a stationary solution of the Euler equation for all of these values of $\tilde{\theta},$ as shown in the figure. However, for values of $\tilde{\theta}$ only moderately larger than the critical value (such as $\tilde{\theta} = 0.20$), this is not the only local minimum, and the global minimum is instead at an interior choice for λ_t; this value results in a path $\{\eta_t\}$ that converges to a different stationary value for η_∞, on the lower branch of the correspondence (as shown for example by the blue dot for $\tilde{\theta} = 0.20$). Yet for values of $\tilde{\theta}$ that exceed a second critical value just above 0.28, the global minimum shifts from the interior minimum to the local minimum at $\eta_{t+1} = \eta_0$. For all values beyond this point, the optimal memory structure involves $\lambda_t = 0$ for all t, so that $\eta_\infty = \eta_0$ (as shown by the blue dots on the upper branch of the correspondence).

Thus while the locus of fixed points $\eta_\infty(\lambda)$ is the same in Figures 1 and 3, all points on this locus represent possible long-run memory structures (attainable through an appropriate choice of $\bar{\lambda}$) in the case of a fixed upper bound on mutual information, but not all of them are always attainable in the case of a linear memory cost function. In the case $K = 1$, the two sets of long-run solutions are identical; but in the case $K = 10,$ there is a range

[44] See Appendix F.3 for a numerical example.

of values for η_∞ that are associated with particular (relatively low) values of $\bar{\lambda}$ but do not correspond to any possible value of $\tilde{\theta}$.[45]

4.3 Stationary fluctuations in the long run

Because our model implies that a DM does not learn the true value of μ with certainty even in the long run, despite an arbitrarily long sequence of observations of the external state, over which time the coefficients of the data-generating process (3.14) are assumed not to change, it follows that the DM's forecasts can be quite different from rational-expectations forecasts — that is, the forecasts of an ideal statistician who knows the true coefficient values. From the standpoint of an observer who is able to determine the true process, the forecasts of the DM with limited memory will appear to be systematically biased. The biases in the DM's forecasts will furthermore fluctuate over time, in response both to variations in the external state (to which the DM reacts differently than someone with rational expectations would) and to noise in the evolution of the memory state.

We obtain a particularly simple characterization of the systematic pattern of forecast biases if we consider the long run — the predictions of the equations in the previous two sections in the case of very large values of t, so that η_t has converged to the constant value η_∞, λ_t has converged to λ_∞, and so on. In this case, our model, like the model of "natural expectations" of Fuster *et al.* (2010, 2011), predicts a stationary pattern of forecast biases that do not reflect incomplete adjustment to a new environment.

In the long run, equations (3.14), (3.19), and (3.25) become a system of linear equations with constant coefficients and Gaussian innovation terms, describing the evolution

[45] We can show analytically that the continuous relationship shown in the left panel of Figure 5 occurs for all $K \leq 1$ when $\rho = 0$, while the backward-bending correspondence and consequent discontinuous relationship between $\tilde{\theta}$ and η_∞ occurs for all $K > 1$. See Appendix F.4 for further explanation.

of the DM's cognitive state. This system of equations can be reduced to a VAR(1) system

$$\tilde{s}_{t+1} \;=\; f\mu \;+\; F\tilde{s}_t \;+\; u_{t+1}, \qquad u_{t+1} \;\sim\; N(0,\, \Sigma_u) \tag{4.46}$$

where

$$\tilde{s}_t \;\equiv\; \begin{bmatrix} \tilde{m}_t \\ \\ y_t \end{bmatrix}, \qquad u_{t+1} \;\equiv\; \begin{bmatrix} \tilde{\omega}_{t+1} \\ \\ \epsilon_{y,t+1} \end{bmatrix},$$

and f, F and Σ_u are a 2-vector and two 2×2 matrices of constant coefficients respectively. In this vector system, the first equation is obtained by substituting (3.19) into (3.37), while the second equation is given by (3.14).

The matrix F furthermore has an upper-triangular form, while Σ_u is diagonal. We show in the appendix that the eigenvalues of the matrix F are ρ and ρ_m.[46] We further show that $0 < \rho_m < 1$, so that both y_t and $\tilde{m}_t$ exhibit stationary fluctuations around well-defined long-run average values which depend linearly on μ. The two independent exogenous sources of variation in this system are the innovations $\epsilon_{y,t+1}$ in the external state and the memory noise innovations $\tilde{\omega}_{t+1}$.

The DM's optimal estimate of μ at each point in time, $\hat{\mu}_t$, as well as her optimal forecast of the external state at any horizon $\tau > t$,

$$\hat{y}_{\tau|t} \;=\; \mathrm{E}[y_\tau \,|\tilde{m}_t, y_t] \;=\; (1-\rho^{\tau-t})\hat{\mu}_t \;+\; \rho^{\tau-t} y_t, \tag{4.47}$$

will then be linear functions of the elements of $\tilde{s}_t$, with coefficients that are also time-invariant. We thus obtain a stationary multivariate Gaussian distribution for any number of leads and lags of the external state, the DM's memory state, and the DM's estimates and forecasts. This allows us to analyze not only the extent to which the DM's forecasts should differ from rational-expectations forecasts, but the correlation that one should observe

[46] See Appendix G.1 for the derivation.

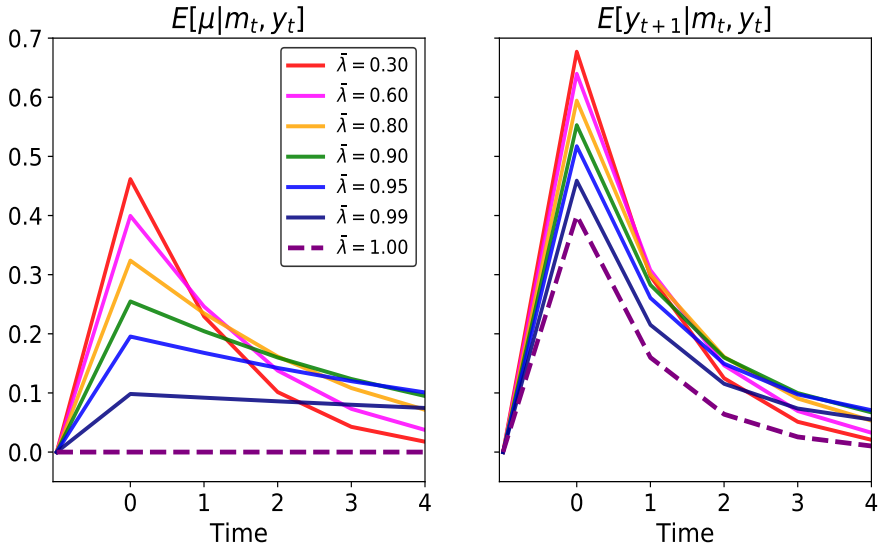


Figure 1.6: Impulse responses

Impulse responses of the DM's estimate of μ (left panel) and one-period-ahead forecast of the state (right panel) to a unit positive innovation in the observed value of y_t at the time marked as "time = 0" on the horizontal axis. Responses are plotted for alternative values of the information bound $\bar{\lambda}$, in the case that $K = 1, \rho = 0.4$.

between the bias in the DM's forecasts and other observable variables.

In particular, the biases in the DM's forecasts will be correlated with the evolution of the external state. An unexpectedly high observed value for y_t will be interpreted (because of the DM's uncertainty about μ) as implying a higher optimal estimate of μ, and this increase in the DM's estimate of μ will furthermore persist, decaying only gradually in subsequent periods. This is illustrated in the left panel of Figure 6, which shows the impulse response function for $\hat{\mu}_\tau$ to a unit positive innovation in the value of y_t. The response is plotted for a variety of alternative values for the information bound $\bar{\lambda}$, in the case that $K = 1$ and $\rho = 0.4$.[47]

In the case that $\bar{\lambda} = 1$ (perfect memory), the value of μ is learned with perfect precision, and as a consequence there is no effect (in the long run, depicted here) of fluctuations in y_t on the DM's estimate of μ. (The Kalman gain γ_1 has a long-run value of zero in this case.)

[47] See Appendix G.1 for illustration of how this figure would change under alternative assumptions about the degree of persistence of the fluctuations in the external state.

Instead, for values of $\bar{\lambda} < 1$, a higher observed value of y_t leads the DM to increase her estimate $\hat{\mu}_t$ (the Kalman gain is positive). The estimate $\hat{\mu}_\tau$ remains higher (on average) in subsequent periods as well. The memory state $\tilde{m}_{t+1}$ carried into the period following the innovation is a noisy record of $\hat{\mu}_t$, and hence is higher because of the increase in y_t; this increases the average value of the estimate $\hat{\mu}_{t+1}$, which increases the average value of the memory state $\tilde{m}_{t+2}$, and so on. The tighter the memory constraint (the lower the value of $\bar{\lambda}$), the greater the effect of the innovation in y_t on $\hat{\mu}_t$, because the DM is more uncertain about the value of μ before observing y_t; however, the effect on the DM's estimate of μ is also more transient the lower the value of $\bar{\lambda}$, because less information is retained from one period to the next about past cognitive states.

These effects on the DM's optimal estimate of μ then feed into her optimal forecast of the external state at any future horizon τ, because of (4.47). As an illustration, the right panel of Figure 6 shows the impulse response of the one-quarter-ahead forecast $\hat{y}_{\tau+1|\tau}$ to a unit positive innovation in y_t, using the same conventions as in the left panel.[48] When $\rho > 0$, the rational-expectations forecast (corresponding to $\bar{\lambda} = 1$ in the figure) is itself increased by a positive innovation in y_t (by an amount equal to fraction ρ of the innovation), and the increase in the forecast is furthermore persistent (decaying back to its original level at a rate proportional to $\rho^{\tau-t}$). But when $\bar{\lambda} < 1$, the forecast is increased by even more, owing to the fact that the higher observation of y_t increases the DM's estimate of μ as well. This additional effect on the forecast is initially larger the smaller is $\bar{\lambda}$; but a smaller $\bar{\lambda}$ (tighter memory constraint) also causes the additional effect to die out more rapidly, since its propagation can only be through the DM's memory of her previous judgment about the value of μ.

Thus our model predicts that forecasts of the future value of a variable will over-react to news about the current value of that variable (assuming, as is often the case with economic time series, that the variable in question exhibits positive serial correlation). Positive serial

[48] The corresponding impulse responses for alternative values of ρ are again shown in Appendix G.1.

correlation means that a higher current observation should increase somewhat one's forecast of the variable's future value, even under rational expectations; but imperfect memory results in a larger increase in the forecast than is consistent with rational expectations. The model also predicts that biases of this kind will persist for some time. Once a situation occurs that leads the DM to over-estimate the future level of some time series, the DM will as a consequence continue (on average) to over-estimate the future level of that variable for several more quarters.

4.4 "Recency bias" in expectation formation

One type of systematic difference between observed expectations and those of a perfect Bayesian decision maker that has often been reported is "recency bias" (e.g., Malmendier and Nagel, 2016; Malmendier *et al.,* 2020) — a tendency for expectations to be influenced more by more recent observations, even when in principle, observations of a given time series at earlier dates should be equally relevant as a basis for inference. As we have already previewed in section 1.3, our model predicts that such a bias should exist, as a consequence of optimal adaptation to limited memory precision (or to the cost of maintaining a more precise memory). Observations of the external state farther in the past are recalled with more noise, and as a consequence are given less weight in estimating parameters of the data generating process than would be optimal in the case of a perfect memory of past data.

The system (4.46) implies that, in the case that data have been generated in accordance with this law of motion for a sufficiently long time, we can express the value of the memory state $\tilde{m}_{t+1}$ as a function of the sequence of external states $\{y_\tau\}$ for $\tau \leq t$ and the sequence of memory noise realizations $\{\tilde{\omega}_{\tau+1}\}$ for $\tau \leq t$:

$$\tilde{m}_{t+1} \;=\; F_{12} \cdot \sum_{j=0}^{\infty} (\rho_m)^j y_{t-j} \;+\; \tilde{\omega}_{t+1}^{sum}, \tag{4.48}$$

where F_{12} is the $(1,2)$ element of the matrix F in (4.46) and

$$\tilde{\omega}_{t+1}^{sum} \equiv \sum_{j=0}^{\infty} (\rho_m)^j \tilde{\omega}_{t+1-j} \tag{4.49}$$

is a serially correlated Gaussian noise term.[49]

Equation (3.19) implies that a DM's estimate of the unknown mean μ of the external state is given by a linear relation of the form

$$\hat{\mu}_t = \xi \tilde{m}_t + \gamma_1 y_t, \tag{4.50}$$

where the coefficient $\xi > 0$ is defined in the appendix. Using (4.48) to substitute for the memory state in this expression, we see that we can write the estimate in the form

$$\hat{\mu}_t = \sum_{j=0}^{\infty} \alpha_j y_{t-j} + \xi \tilde{\omega}_t^{sum}, \tag{4.51}$$

where the weights $\{\alpha_j\}$ are all positive, and the weights for $j \geq 1$ decrease exponentially: $\alpha_j = \alpha_1 (\rho_m)^{j-1}$.

The forecasts specified by (4.47) using this value for $\hat{\mu}_t$ are similar to those implied by a model of least-squares learning (Evans and Honkapohja, 2001) in which the DM is assumed to know that the variable's law of motion is of the form (3.14); the value of the coefficient ρ is assumed to be known while μ must be estimated; and the unknown coefficient is estimated using a "constant-gain" estimator.[50] The biases in forecasts predicted by our model will therefore have important similarities to those of a model of constant-gain learning, of the kind included in estimated macroeconomic models by authors such as Milani (2007, 2014) and Slobodyan and Wouters (2012).

[49] This is a stationary random process with a finite unconditional variance, since $0 < \rho_m < 1$ as shown in Appendix G.1.

[50] The differences between (4.51) and a standard constant-gain estimate of the mean of a series are the fact that the coefficient α_0 is differently specified, and the presence of the Gaussian error term. See further discussion in section 5.1.2 below.

We provide, however, a justification for the declining weight on observations farther in the past, as a consequence of optimal forecasting based on an imperfect memory, and furthermore endogenize the nature of that memory. The fact that our model predicts decreasing weights on observations made farther in the past is a notable difference between our model and the one proposed by Afrouzi *et al.* (2020), as we discuss further in section 5.2.2.

5 Experimental Evidence

We have shown that our model provides an explanation for important qualitative features of observed subjective expectations. Here we briefly discuss the model's quantitative fit with data on subjective expectations from the laboratory experiment of Afrouzi *et al.* (2020). We focus on this particular evidence for a quantitative test of our model, because it involves forecasts of a stationary AR(1) process, and in that sense matches exactly the problem assumed in our theoretical analysis above. A laboratory experiment also has the advantage over field studies of allowing us to be sure exactly what the true data-generating process is, and exactly what information is available to decision makers at each point in time (though of course questions remain about how the situation is understood by the experimental subjects, and what they pay attention to).

As noted in the introduction, Afrouzi *et al.* (2020) conduct a laboratory experiment in which subjects observe successive realizations of an AR(1) process, and forecast what the next realizations should be. They find that subjects' reported expectations over-react to innovations in this process, as predicted by our model (as well as the related model of noisy memory that they discuss). They give particular emphasis to a measure of over-reaction in which a subject's forecast $\hat{y}_{t+h|t}$ (where h is the number of realizations in advance for which the forecast is solicited in trial t) is regressed on the realization of the variable just

before the forecast is solicited:

$$\hat{y}_{t+h|t} \;=\; \alpha_h^{subj} \;+\; \rho_h^{subj} y_t \;+\; v_t. \tag{5.52}$$

A separate regression (with coefficients α_h, ρ_h^{subj}) can be estimated for each of several horizons h. Afrouzi *et al.* are interested in the difference between the "subjective degree of persistence" measured by the estimated coefficient ρ_h^{subj} and the corresponding coefficient ρ_h in a regression using actual outcomes:

$$y_{t+h} \;=\; \alpha_h \;+\; \rho_h y_t \;+\; u_{t+h}. \tag{5.53}$$

The authors measure the degree of over-reaction of expectations to news by the extent to which ρ_h^{subj} is larger than ρ^h. Note that this is an example of a test of the predictability of forecast errors, since the coefficient of a regression of the forecast error $y_{t+h} - \hat{y}_{t+h|t}$ on y_t will equal $\rho^h - \rho_h^{subj}$.

We can investigate what our model of expectation formation on the basis of an imperfect memory implies about the relationship between ρ_h^{subj} and ρ_h in the case of a stationary AR(1) process. Here we consider the predicted values of the regression coefficients in the long run, as the length of the time series used to estimate them goes to infinity. The law of motion (3.14) implies that for any horizon $h \geq 1$, the joint distribution of y_t and y_{t+h} (conditional on the value of μ) will be bivariate Gaussian, with

$$\mathrm{E}[y_{t+h} \,|\mu, y_t] \;=\; (1-\rho^h)\mu \;+\; \rho^h y_t.$$

Hence with a sufficiently long series of observations, the coefficients in a regression of the form (5.53) should approach the asymptotic values

$$\alpha_h \;=\; (1-\rho^h)\mu, \qquad \rho_h \;=\; \rho^h.$$

(Here we assume that the regression uses an arbitrarily long sequence of realizations of a process for which there is a single, unchanging value of μ.)

Equation (4.47) implies that subjective forecasts should be given by

$$\hat{y}_{t+h|t} \;=\; (1-\rho^h)\hat{\mu}_t \;+\; \rho^h y_t,$$

so that the predicted coefficient ρ_h^{subj} in regression (5.52) will equal

$$\rho_h^{subj} \;=\; (1-\rho^h)\beta_{\hat{\mu}|y} \;+\; \rho^h \;=\; (1-\rho_h)\beta_{\hat{\mu}|y} \;+\; \rho_h, \tag{5.54}$$

where $\beta_{\hat{\mu}|y}$ is the coefficient in a regression of $\hat{\mu}_t$ on y_t,

$$\beta_{\hat{\mu}|y} \;=\; \frac{\text{cov}[\hat{\mu}_t, y_t\,|\mu]}{\text{var}[y_t\,|\mu]} \;=\; \frac{\text{cov}[\hat{\mu}_t, y_t\,|\mu]}{\sigma_y^2}.$$

We show in the appendix how to calculate this coefficient as a function of the model parameters.[51]

Importantly, our numerical solutions indicate that $\hat{\mu}_t$ and y_t are always positively correlated (conditional on μ). This is because a positive innovation in the external state y_t raises (or at least never lowers) the expected value of y_τ for all $\tau \geq t$, and at the same time also raises the expected value of $\hat{\mu}_\tau$ for all $\tau \geq t$ (as illustrated in Figure 6 and similar figures in the appendix). Since the memory noise has no effect on the evolution of the external state, there are no shocks that move $\hat{\mu}_t$ and y_t in opposite directions, while some (at least the innovation ϵ_{yt}) move both of them in the same direction. But given that $\beta_{\hat{\mu}|y} > 0$, equation (5.54) implies that $\rho_h^{subj} > \rho_h$; that is, our model implies over-reaction of the kind exhibited by the forecasts of the subjects of Afrouzi *et al.*

Equation (5.54) also implies that for fixed values of the model parameters other than ρ, the over-reaction measure $\rho_h^{subj} - \rho_h$ converges to zero as $\rho \to 1$, for any forecast horizon

[51] See Appendix G.3 for details.

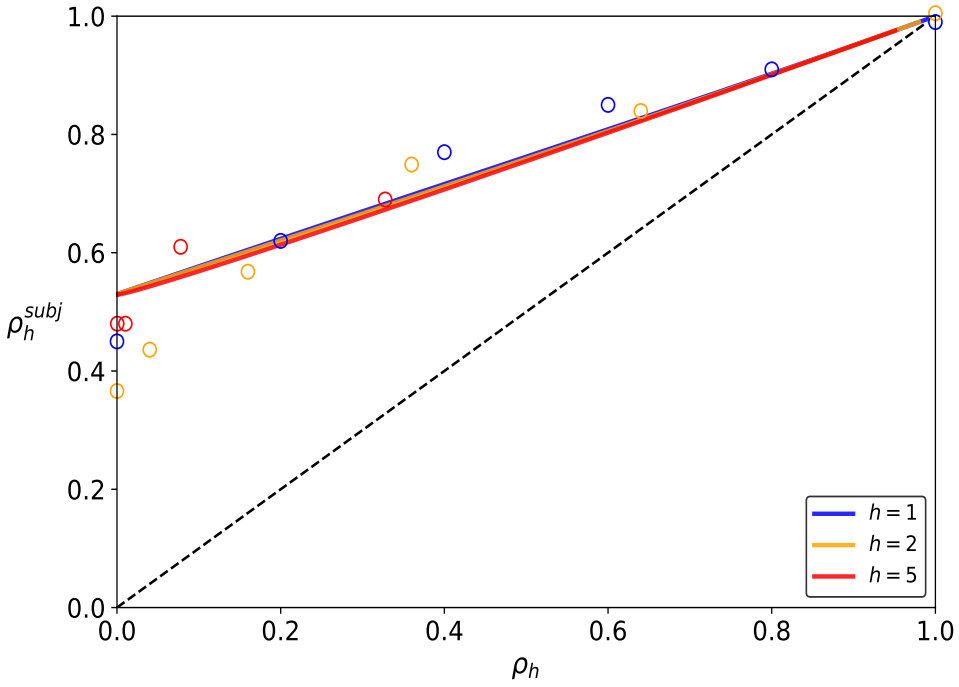


Figure 1.7: Two regression coefficients

Comparison of the values for the regression coefficients ρ_h and ρ_h^{subj} for different values of ρ and h. (The figure is shown for the case $K = 1, \bar{\lambda} = 0.3$.) The diagonal line indicates the prediction of the rational-expectations hypothesis.

h.[52] This is also approximately true of the regression coefficients reported by Afrouzi *et al.* (see their Figures 2B, 5A, and 5B). Indeed, these authors stress the finding that in their data, the discrepancy $\rho_h^{subj} - \rho_h$ is much larger when ρ_h is relatively small (either because ρ is small, or because ρ is well below one and h is large). This is also true in numerical solutions of our model as indicated in Figure 7.

One of the more striking features of the regressions reported by Afrouzi *et al.* is that ρ_h^{subj} is well approximated by an increasing function of ρ_h, with approximately the same functional relationship regardless of whether the variation in ρ_h occurs as a result of vari-

[52] This prediction depends on $\beta_{\hat{\mu}|y}$ remaining bounded as ρ approaches 1. This is the case in our numerical solutions, both when $\bar{\lambda}$ is held constant as ρ is varied (as in Figure 2) and when $\tilde{\theta}$ is held constant as ρ is varied (as in Figure 4).

ation in ρ or variation in h.[53] The relationship $\rho^{subj}(\rho)$ is furthermore an upward-sloping one, with a slope much less than one, starting well above the diagonal for low values of ρ and approaching the diagonal as $\rho \to 1$. (See the plot of their regression coefficients in Figure 7.[54]) While our model does not imply that a functional relationship of that kind should hold precisely, it is worth noting that to the extent that the value of $\beta_{\hat{\mu}|y}$ remains approximately the same as one varies ρ, (5.54) implies that the value of ρ_h^{subj} should be nearly the same for all pairs (ρ, h) that imply the same value of ρ_h. Perhaps more to the point, our model can be parameterized so that it simultaneously fits the experimental evidence for each of the three different horizons for which forecasts are solicited in the experiment of Afrouzi *et al.*

Figure 7 plots the predicted value of ρ_h^{subj} against the value of ρ_h, for each of several different horizons h, each represented by a distinct curve; the curves are shown for the case in which $K = 1$ and $\bar{\lambda} = 0.3$. Along each curve, the variation in ρ_h is due purely to variation in ρ. (The fact that $\bar{\lambda}$ is fixed despite variation in ρ means that we assume a fixed upper bound on the mutual information, as in section 1, rather than a convex cost function.) The horizons used are $h = 1, 2$ and 5, as these are the horizons for which Afrouzi *et al.* elicit forecasts from their subjects; the regression coefficients that they estimate for various combinations of ρ and h are indicated by the circles in the figure (with colors indicating the horizon h).

The three curves are not exactly the same, since in our model $\beta_{\hat{\mu}|y}$ is a function of ρ (but the same for all values of h), rather than being a function only of ρ_h. Nonetheless, for the parameterization chosen here, $\beta_{\hat{\mu}|y}$ is nearly constant as ρ is varied; as a consequence, the relationship between ρ_h and ρ_h^{subj} predicted by (5.54) is close to a linear one, and is nearly the same for all values of h. Our model therefore provides quite a good account of the effects of variation in either ρ or h on the value of ρ_h^{subj}, as indicated by the fact that

[53] This was shown in an earlier version of the paper now circulated as Afrouzi *et al.* (2020), though this figure is omitted from their most recent draft.

[54] The data plotted here are based on Figures 2B, 5A, and 5B of Afrouzi *et al.* (2020).

none of the circles in Figure 7 are far from the corresponding curve.

There is also evidence of over-reaction to news in the forecasts of macroeconomic and financial variables by professional forecasters, as discussed by Bordalo *et al.* (2020). A satisfactory quantitative account of the predictable forecast errors observed in these forecasts requires an extension of the model presented here, as discussed by Sung (2022). While the more complex model in that paper involves additional information frictions, as addition to allowing for more complex dynamics of the variables that are forecasted, noisy memory of the kind modeled here remains crucial for explaining the observed patterns. And while information frictions of the kind proposed by Coibion and Gorodnichenko (2012, 2015) are also important, Sung finds that quantitative estimates of the size of those frictions are significantly biased by failing to take account of the effects of noisy memory.

6 Related Models

Here we compare our model to alternative models of belief formation that make at least somewhat similar predictions, most notably with regard to the possibility of over-reaction to recent news. We show how our model has important formal similarities to some of these others, and clarify the ways in which it differs from them.

6.1 Alternative Explanations for Over-Reaction

We begin by reviewing possible explanations for over-reaction to news that do not rely upon imperfect memory. To simplify the discussion, we here consider only possible explanations for a pattern of over-reaction that would continue to be observed even after an arbitrarily long sequence of observations (rather than discussing transitory dynamics that depend on the DM having insufficient experience with a given context).

Reactions to News when the Mean is Understood to Drift

In our model, over-reaction of forecasts to new observations of the variable y_t reflect revisions of the DM's estimate of the mean of the stochastic process $\{y_t\}$, even though the mean μ is assumed to be constant over time; failure of the DM to learn the exact value of μ, even in the long run, depends on imperfect memory. However, there would be perpetual revision of beliefs about the mean, even with perfect memory (and perfect Bayesian inference) in a world where the mean is (correctly) understood to evolve stochastically over time. In this case, observations farther in the past would be of progressively less relevance to the DM's current estimate of the mean, even with perfect memory.

Such a model can predict forecast dynamics similar (though not identical) to those in our model. As a simple example, suppose that $y_t = \mu_t + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma_y^2)$ represents an i.i.d. deviation from the time-varying mean μ_t; and suppose that the mean evolves according to an AR(1) process,

$$\mu_t = \phi\mu_{t-1} + \nu_t,$$

where $0 < \phi < 1$ and $\nu_t \sim N(0, (1-\phi^2)\Omega)$ is another i.i.d. process. Note that this specification implies that the unconditional prior distribution for the mean is given by $N(0, \Omega)$, just as in our model.[55] Let us consider the evolution of the beliefs of a perfect Bayesian DM in such an environment, who observes y_t with perfect precision each period, starting from prior beliefs about μ_0 (before y_0 is observed) corresponding to the unconditional prior.

The posterior distribution for the value of μ_t, just before the observation of y_t, will be a Gaussian distribution $N(\hat{\mu}_{t|t-1}, \hat{\sigma}^2_{t|t-1})$. The posterior after observing y_t will be another Gaussian distribution $N(\hat{\mu}_{t|t}, \hat{\sigma}^2_{t|t})$. The mean and variance of this distribution are given by the same Kalman-filter formulas (2.6)–(2.8) as above.[56] In the case of perfect memory,

[55] The difference is that in the drifting-mean model, we no longer assume that a value of the mean is drawn from this distribution and then remains constant forever after. Our specification in the previous sections can be regarded as the $\phi \to 1$ limit of this prior.

[56] Here we must substitute $\hat{\mu}_{t|t-1}$ for the prior mean $\bar{m}_t$ in equation (2.6), and $\hat{\sigma}^2_{t|t-1}$ for the prior variance Σ^{μ}_t. Similarly $\hat{\mu}_{t|t}$ and $\hat{\sigma}^2_{t|t}$ correspond to the variables called simply $\hat{\mu}_t$ and $\hat{\sigma}^2_t$ in the previous equations.

these posterior beliefs about μ_t imply a posterior distribution for the value of μ_{t+1} that is also Gaussian, with mean and variance

$$\hat{\mu}_{t+1|t} \;=\; \phi\hat{\mu}_{t|t}, \tag{6.55}$$

$$\hat{\sigma}^2_{t+1|t} \;=\; \hat{\sigma}^2_{t|t} \;+\; (1-\phi^2)(\Omega - \hat{\sigma}^2_{t|t}). \tag{6.56}$$

Comparison of equation (6.56) with the corresponding equation (2.11) for the dynamics of posterior uncertainty in our noisy-memory model reveals that the degree of uncertainty, after any particular number of observations of y_t, is the same in both models in the case that $\phi^2 = \bar{\lambda}$. The same path for uncertainty about the mean then implies the same path for the Kalman gain γ_t, given by (2.7). Hence the perfect-Bayesian model with a stochastic mean is equally capable of explaining why a DM's estimate of the mean should continue to be influenced by recent observations, even after a long sequence of observations; the predictions of the two models about this are identical, if the parameter ϕ is chosen appropriately.

However, this does not mean that the two models are observationally equivalent. Equations (6.55)–(6.56) together with the Kalman-filter equation (2.6) imply that after any finite sequence of observations the Bayesian estimate of the mean will be given by a solution of the form (2.12), but with the weights $\{\alpha_j\}$ given by

$$\alpha_{j,t} \;=\; \phi^j \gamma_{t-j} \Pi_{i=1}^{j} (1-\gamma_{t-j+i})$$

and the weights $\beta_{j,t} = 0$. If we assume that $\phi^2 = \bar{\lambda}$, so that the Kalman gains implied by both models are the same, the weights on past observations $\{y_\tau\}$ will equal

$$\alpha_{j,t} \;=\; \bar{\lambda}^{j/2} \gamma_{t-j} \Pi_{i=1}^{j} (1-\gamma_{t-j+i}).$$

The weights decay exponentially, like the weights (2.13) implied by the noisy-memory

model; but they do not decay at the same rate. (The weights decay more rapidly as j increases in the case of the noisy-memory model; hence the weights are smaller for all $j \geq 1$ in that case.)

Another difference between the two models is that the perfect-Bayesian model implies that there should be a tight relationship between the degree of persistence of the series $\{\mu_t\}$ — and hence the autocorrelation of the observed series $\{y_t\}$ — and the coefficients (such as the Kalman gain γ_t) that describe the dynamics of beliefs about the mean. In the noisy-memory model, the coefficient $\bar{\lambda}$ that determines the size of the Kalman gain and the intrinsic persistence of the belief state can be specified independently of the time-series properties of the process $\{y_t\}$. This flexibility is important for accounting for observed beliefs. Bayesian models of subjective forecasts often have to posit a DM with an apparent prior belief that an unknown state fluctuates more than is actually the case.[57] The noisy-memory model can account for such findings without having to suppose that people fail to learn the correct statistics of their environment. The example just presented shows that noisy memory ($\bar{\lambda} < 1$) can result in belief dynamics similar to those of a Bayesian model in which the DM's prior assumes that the mean μ_t is less persistent than it really is (the prior assumes that $\phi < 1$ when actually μ never changes).

Finally, the perfect-Bayesian model implies that the DM's estimate $\hat{\mu}_{t|t}$ at any time (and hence their forecasts) will be a deterministic function of the sequence of values $(y_0, \ldots, y_t)$ that have been observed. It follows from this that all forecasters who observe the same series should have identical forecasts, and that the variation over time in their forecasts can be fully accounted for by the variation in the values that have been observed. The noisy-memory model instead implies that each DM's beliefs (and hence their forecasts) are affected by memory noise ($\beta_{j,t} \neq 0$); this implies both that forecasts are not perfectly predictable from the past history of the series being forecasted, and that they should differ across forecasters. This is an attractive feature of the noisy-memory model, since observed

[57] See, for example, Yu and Cohen (2009).

forecasts have both of these properties.[58]

Constant-Gain Learning

One also obtains a prediction of perpetual learning, and hence continued over-reaction to news even after an arbitrarily long sequence of observations, in a model where the DM is assumed to estimate the value of the parameter μ using a "constant-gain" variant of least-squares learning (Evans and Honkapohja, 2001, sec. 7.4). Constant-gain (CG) algorithms effectively put an exponentially decreasing weight on observations farther in the past; for example, an unknown mean is estimated by a linear estimator of the form

$$\hat{\mu}_t \;=\; \sum_{j=0}^{t} \gamma(1-\gamma)^j y_{t-j} \;+\; (1-\gamma)^{t+1}\hat{\mu}_{-1},$$

where $0 < \gamma < 1$ is the constant "gain factor" and $\hat{\mu}_{-1}$ is an initial condition (representing the state of belief before y_0 is observed). If we set $\hat{\mu}_{-1} = 0$ (in accordance with the prior assumed in our noisy-memory model), this is similar to the kind of estimate of the unknown mean implied by the perfect-Bayesian model in the case of a drifting mean.[59] Moreover, in the CG algorithm, the value of γ can be specified independently of the dynamics of the process $\{y_t\}$ that is forecasted.[60]

To the extent that a model of CG learning is considered to be empirically realistic, however, a question arises as to what determines the value of the gain parameter. In the adaptive control literature, such algorithms are proposed as a way of dealing with drift in the values of parameters to be estimated; the appropriate value of the gain parameter

[58] See Sung (2022) for discussion of the difference between individual professional forecasters' forecasts and the consensus forecast, in the case of a variety of macroeconomic variables.

[59] Note that if we consider a limiting case in which $\phi \to 1$ while $(1-\phi^2)\Omega \to \sigma_\nu^2 > 0$, then as $t \to \infty$ the solution (2.12) for the perfect-Bayesian model approaches one in which $\gamma_{t-j} \to \bar{\gamma}$, a constant value between 0 and 1, for all j. In this case the dynamics of the mean estimate implied by the perfect-Bayesian model are exactly those of a constant-gain mean estimate with a gain factor of $\bar{\gamma}$.

[60] In empirical applications (e.g., Milani, 2007, 2014; Slobodyan and Wouters, 2012), the gain parameter and the parameters specifying the persistence of the exogenous states are treated as independent free parameters to be estimated.

should thus depend on one's prior regarding the degree of volatility of the parameters to be estimated, as in our discussion above of Bayesian inference when the mean drifts. But once again, the gain parameters that are found to best fit expectations data do not seem to correspond to ones that would be optimal given the degree of structural change in the forecasted time series.[61] Alternatively, authors such as Malmendier and Nagel (2016) propose that aggregate dynamics similar to those predicted by a model of CG learning can result from aggregation of the decisions of people of different ages, who each form beliefs on the basis of their personal experience (and hence on the basis of samples extending different distances into the past).[62] But also under this explanation for CG learning, the predicted gain parameter should depend on other features of the model, that may not justify a gain parameter as large as the one required to explain the observed degree of over-reaction to news.[63] Our model provides an alternative foundation for belief dynamics similar to those implied by a CG algorithm, in which a substantial gain parameter can exist even when the value of the mean remains constant (or nearly constant) over long periods of time, and even when forecasts have long personal histories of observations.

Forecasts Based on an Incorrect Model

A longstanding explanation for systematic over-reaction to news is the hypothesis that people form their forecasts on the basis of an incorrect statistical model — for example, under an assumption that the fluctuations in $\{y_t\}$ are more persistent than is actually the case. Explanations of this kind have continued to be prominent in the recent literature (e.g., Angeletos *et al.,* 2021), but they raise the question: why should people persist in mis-estimating the dynamics?

[61] See, e.g., Branch and Evans (2006) and Berardi and Galimberti (2017).

[62] For additional examples, see Nakov and Nuño (2015), Schraeder (2016), Collin-Dufresne *et al.* (2017), Ehling *et al.* (2018), and Malmendier *et al.* (2020).

[63] Thus Malmendier *et al.* (2020) posit an exponentially decaying influence of earlier experiences on a given DM's expectations, even among the events that have occurred during their lifetime, rather than relying upon demographics alone to account for the rate at which past events cease to influence current market pricing.

Fuster *et al.* (2010, 2011) offer one answer: people's forecasts are optimal, given their estimated model of the dynamics, and their estimated model is the one that best fits the autocorrelation function of the actual series, within some parameteric family of possible models (that need not include the true data-generating process). Their hypothesis of "natural expectations" assumes that the class of statistical models considered is that of all possible AR(k) models, for some fixed bound on k.[64] The authors argue that actual time series often involve long-horizon dependencies, and show that in this case (say, an AR(40) process forecasted by people who consider models with no more than 10 lags), long-horizon forecasts using the best-fitting AR(k) model can significantly over-react to recent trends in the data.

This proposal, however, remains subject to several objections. Why should the restriction to models of the data with a fixed upper bound on k be maintained, even when the available sequence of observations with which to estimate the model becomes unboundedly long? Moreover, even if one grants that a constraint on model complexity requires that no more than some finite number of explanatory variables be stored and used as a basis for forecasts, why must the possible explanatory variables correspond only to the last k observations of the series? In the kind of example in which Fuster *et al.* argue that their proposal predicts over-reaction, more accurate long-horizon forecasts would be possible if the forecast were conditioned on a long moving average of observations, rather than only recent observations; yet tracking a small number of moving averages would seem no more complex than always having access to the last k observations. And above all, the Fuster *et al.* explanation implies that over-reaction should only be observed in the case of variables that are not well-described by an AR(k) process of low enough order. Yet as discussed above, Afrouzi *et al.* (2020) find significant over-reaction in an experiment in which the true data-generating process is an AR(1) process; and in fact, they find the most severe degree of over-reaction when the process to be forecasted is white noise.

[64] More general versions of this hypothesis are considered in the more recent work of Molavi (2022).

Like the hypothesis of "natural expectations," our model assumes that forecasts are optimal, among those forecasting rules in which the forecast is based on only a limited summary of past history; but the way in which we model the limit on the complexity of possible representations of past data is different. Our approach does not impose any *a priori* restriction on either the dimensionality of the memory state or the number of past observations that can be (imperfectly) represented by the memory state. And the form of complexity limit that we assume has the advantage of implying forecasting bias (and more specifically over-reaction) even when the true dynamics are very simple — indeed, even when the true dynamics are white noise (and are recognized by the DM to be white noise).

6.2 Alternative Models of Imprecise Memory

We are also not the only authors to have proposed that expectational biases may result from forecasts being based on imperfect memory of past observations. Here we briefly discuss similarities and differences of alternative proposals with our own approach.

Models of Quasi-Bayesian Belief Updating

Nagel and Xu (2022) propose that a variety of asset-pricing anomalies can be explained by the biases in expectations regarding future asset returns implied by a particular type of departure from perfect Bayesian inference from observed past returns, which they call a model of "fading memory." As in this paper, they consider a situation in which a DM (an investor) must infer the mean μ of a process $\{y_t\}$, based on past observations of this process; and (as in the simple case analyzed in section 1) they assume that the process is i.i.d. (and known to be), and that the only unknown parameter of the distribution is μ. Given a prior $p(\mu)$ over possible values of μ, and a likelihood $p(y\,|\mu)$ for the observation of y_t in any period conditional on the unknown mean, the Bayesian posterior distribution

conditional on a finite sequence of observations $\mathbf{y} = (y_{t_0}, \ldots, y_{t-2}, y_{t-1}, y_t)$ is given by

$$p(\mu \,|\boldsymbol{y}) \;\sim\; p(\mu) \prod_{j=0}^{t-t_0} p(y_{t-j} \,|\mu).$$

The Nagel-Xu model of "fading memory" instead assumes a subjective posterior of the form

$$p(\mu \,|\boldsymbol{y}) \;\sim\; p(\mu) \prod_{j=0}^{t-t_0} p(y_{t-j} \,|\mu)^{(1-\nu)^j}, \tag{6.57}$$

for some small quantity $\nu > 0,$ which indicates the rate at which memory of past observations "fades." (Note that their model reduces to perfect Bayesian inference in the limiting case in which $\nu = 0$.)

The Nagel-Xu model, like ours, is one in which there is perpetual learning: in the limit as $t_0 \to -\infty,$ the posterior distribution (6.57) remains non-degenerate, despite being based on a sample of infinite length. As in our case, the reason is that past observations have a progressively weaker influence on the posterior, the farther they are in the past, and more specifically the influence decreases as an exponential function of the elapsed time. Also as in our case, the Nagel-Xu model implies that one should observe "recency effects." Another important similarity between their approach and ours is that Nagel and Xu model the DM's complete posterior at each point in time, not just the DM's point estimate of μ; and like us, they tie the rate of decay of past information to cognitive limitations, rather than the rate at which the environment is objectively likely to have changed.

Our model differs from that of Nagel and Xu, however, in offering an explicit representation of the imprecise information contained in memory, and then deriving the DM's subjective posterior from (correct) Bayesian conditioning on this imprecise record, rather than directly assuming a particular modification of the Bayesian expression for the posterior beliefs. This is not simply a matter of having failed to provide intermediate steps in the derivation; the subjective beliefs assumed by Nagel and Xu are not correct conditional beliefs, if one were to condition on the information about past observations reflected in

the assumed beliefs (and therefore revealed by the DM's cognitive state, since the subjective posterior must be some function of the cognitive state).[65] Our model also differs from theirs in that it implies that individuals' beliefs involve idiosyncratic cognitive noise; thus our model, unlike that of Nagel and Xu, predicts that investors should have heterogeneous beliefs even if they observe identical information. (This difference is relevant for applications to financial economics, since our model of heterogeneous beliefs on the part of individual investors provides a motive for trading, even when all information about asset fundamentals is public.) In these respects, the predictions of our model are not quantitatively identical to those of the model of Nagel and Xu, despite many similarities.

Prat-Carrabin *et al.* (2021) derive a quasi-Bayesian posterior very similar to the one postulated by Nagel and Xu from a hypothesis of "costly Bayesian inference," in which belief updating after each new piece of evidence arrives is distorted (relative to exact Bayesian updating) so as to reduce the precision of the resulting belief state.[66] This hypothesis is even more closely related to the one that we propose here, insofar as the sensitivity of beliefs to past observations decreases over time as a consequence of a cost of storing a more precise record of the DM's past cognitive state. The model of Prat-Carrabin *et al.* differs from ours in identifying the imprecise memory state with the DM's (distorted) posterior beliefs given the sequence of observations to that point; instead, we distinguish between the DM's cognitive state (which includes the memory state m_t) and the probability beliefs that would optimally be inferred from such a state. Thus again, while there are many similarities between the predictions of their model and ours, the predictions are not identical.

Alternative Models of Noisy Memory

Neligh (2022) proposes a model of decaying memory that is conceptually closer to our own in that, as in this paper, it is assumed that memory can be retrieved only with noise,

[65] See the Appendix, section XX, for detailed discussion.

[66] Prat-Carrabin *et al.* (2022) fits the model to an experimental data set.

and the judgments that are made are optimal (consistent with correct Bayesian inference) subject to being based on the noisy memory state. The difference with our model is in the way that the memory state, and the cost of retrieving a more precise memory, are modeled. As noted above, Neligh assumes an "episodic" memory, in which there is an independent noisy record of each of the past observations y_τ for $0 \leq \tau \leq t-1$; the element of the memory vector corresponding to the observation at time τ is equal to the value of y_τ plus a mean-zero Gaussian noise term, distributed independently of the value of y_τ, and with a variance that depends on the amount of elapsed time. This is a special case of the kind of noisy memory that our framework allows for, but is not the form of memory that is found to be optimal for the decision problem considered in this paper. In addition to imposing the constraint that memory must take this form, Neligh endogenizes the precision of memory in only one respect: the precision with which observation y_τ is initially encoded at date τ is optimized (subject to a cost of greater encoding precision), but given the choice of an initial encoding precision, the precision of the memory that can be retrieved after a time delay is exogenously determined by the amount of time that has elapsed. Our model instead allows the precision with which memory is maintained over time to be endogenously varied.[67]

An important similarity between Neligh's model and ours is that in both models, observations more distant in the past are retrieved with greater noise, because of the way in which noise is cumulatively added as the memory state is maintained over time. This means that both models predict recency effects; and it would be possible to specify the rate of increase of memory noise with the passage of time in Neligh's model in such a way as to make the distribution of $\mathrm{E}[\mu \,|m_t]$ conditional on the sequence of past observations — and hence the conditional distribution of all of the DM's forecasts, in the decision problem considered here — the same as the one predicted by our model. There would remain, however, two important differences between Neligh's model and ours. One is that our model

[67] In addition to considering a different class of possible memory structures, Neligh (2022) addresses largely distinct questions from those analyzed here.

derives its predictions from less special assumptions and involves fewer free parameters; thus in the case that both models were equally consistent with empirical observations like those of Afrouzi *et al.* (2020), our model would provide a more parsimonious explanation. And second, Neligh's model implies a much higher-dimensional memory state than does ours. In the case of the decision problem considered in this paper, this makes no difference, as forecasts depend on memory only through a single scalar summary statistic; but the predictions of the two models would likely be different in the case of more complex decisions.

Like us, Afrouzi *et al.* (2020) propose to explain the biases in their experimental subjects' forecasts using a model of endogenously imprecise memory. However, in their model, all past observations are stored in memory with perfect precision; imprecision enters only when an imperfect representation of the contents of memory is retrieved in order to inform a decision. The nature of the imprecise representation that is used for the decision is optimized subject to a cost of precision, which as in our model is based on mutual information (for them, the mutual information between the complete contents of memory and the imprecise representation). As in our model, the information cost implies that an accurate estimate of the value of μ cannot be made on the basis of memory, even after a very large number of observations. Hence subjects' forecasts (assumed as in our model to be optimal subject to having to be conditioned on an imprecise cognitive state) are based on a precise observation of the current y_t together with an imprecise estimate of μ deriving from an imprecise summary of past observations. This results, as in our model, in a prediction of over-reaction to the most recent observation (that can be observed with greater precision than any past observations are recalled); and the predicted degree of over-reaction is greatest in the case of variables with low persistence (since in this case optimal forecasts are largely determined by the optimal estimate of μ).

Despite these similarities in the predictions of the two models, there is an important difference between the model of noisy memory in Afrouzi *et al.* (2020) and our own.

Their model implies that all past observations are accessible with equal precision when a forecast needs to be made; hence the optimal noisy representation of the past weights past observations to the extent that they are relevant to the current decision, which implies much less "decay" of old observations than in our model. As a simple example, consider the case in which y_t is i.i.d. Then the contents of memory will be distributed independently of the current observation y_t, and the equally-weighted sample mean of the observations $\{y_\tau\}$ for $0 \leq \tau < t$ will be a sufficient statistic for the information about the mean μ that is contained in the previous observations; hence the optimal representation will be a noisy read-out of this sample mean. It follows that any past observation y_τ (for $\tau < t$) should have exactly the same effect on forecasts at time t as any other: there will be no "recency effect" at all, except for the fact that the observation y_t will have a larger effect than any of the observations at dates $\tau < t$. Thus the model of Afrouzi *et al.* provides no explanation for the kind of recency effects that have frequently been documented in the experimental literature (e.g., Hogarth and Einhorn, 1992), as well as in macroeconomic and financial contexts by authors such as Malmendier *et al.* (2020).

7 Conclusion

We have shown that it is possible to characterize the optimal structure of memory, for a class of linear-quadratic-Gaussian forecasting problems, when the cost of a more precise memory is proptional to Shannon's mutual information, and when we assume that the joint distribution of past cognitive states and the memory state is of a multivariate Gaussian form, but with no *a priori* restriction on the dimension of the memory state or the dimensions of past experience that may be more or less precisely recalled. Strikingly, we find that for the class of problems that we consider, the optimal memory structure is necessarily at most one-dimensional. This means that what can be recalled at any time about past observations is simply a noisy recollection of a single summary statistic for past experience. We show how the model parameters determine the law of motion for that sum-

mary statistic, and hence what single dimension of past experience will be (imprecisely) available as an input to the DM's forecasts.

Among the implications of our model, two seem of particularly general interest. First, while our formalism allows for the possibility of an independent noisy record of each past observation (as assumed for example in the model of Neligh, 2022), this is not optimal; instead, the optimal memory structure is one in which only a particular weighted average of past observations can be recalled with noise. And second, this weighted average places much larger weights on recent observations than on ones at earlier dates, even though observations at all dates are equally relevant to inference about the value of the parameter μ, which matters for the DM's decisions. Thus our model provides an explanation for "recency bias" in the influence of past observations on current decisions, unlike the model of endogenous memory precision proposed by Afrouzi *et al.* (2020).

We have shown that our model predicts "over-reaction" of forecasts of an autoregressive process to current realizations of the process, and that the degree of over-reaction should be greater in the case of less persistent time series, as observed in the forecasts of experimental subjects (Afrouzi *et al.,* 2020). The same mechanism provides a potential explanation for the frequent observation of over-reaction to news in survey forecasts of macroeconomic and financial time series (e.g., Bordalo *et al.,* 2020). Sung (2022) extends our model to allow for imprecise awareness of the current external state y_t, in addition to the imprecise awareness of the DM's own past cognitive states modeled in this paper, and shows that with this extension the model can account quantitatively for the predictable errors in professional forecasts of a variety of macro variables. In particular, she shows that the model can simultaneously account for the apparent "under-reaction" of consensus forecasts stressed by Coibion and Gorodnichenko (2012, 2015) and the apparent "over-reaction" stressed by Bordalo *et al.* (2020).

In these applications we have focused on biases observed in people's stated expectations. But we suspect that the expectational biases implied by our model can help to

explain puzzling aspects of market outcomes as well. For example, Bordalo *et al.* (2022) argue that a number of well-known puzzles about the behavior of the aggregate stock market are in fact all consistent with a simple dividend discount model of stock prices, under the hypothesis that market expectations regarding firms' future earnings differ systematically from rational expectations in a particular way, that is furthermore consistent with the biases observed in survey expectations of earnings. They further show that a particular sort of bias in market expectations is needed in order to explain both the biases in survey expectations and the asset pricing anomalies, one very much like the kind of forecast bias predicted by our model.

Briefly, Bordalo *et al.* propose a model in which asset prices at time t are based on market expectations of dividend growth g_{t+h} at various future horizons h. Dividend growth is assumed to be a stationary autoregressive process; market expectations of g_{t+h} differ from rational expectations by an expectional error term $\epsilon_{h,t}$. For any horizon h, $\epsilon_{h,t}$ is assumed to be a stationary, mean-zero autoregressive process, with a substantial degree of persistence; and the innovations in $\epsilon_{h,t}$ are positively correlated with the innovations in g_t, though fluctuations in $\epsilon_{h,t}$ also occur that are uncorrelated with fundamentals. Finally, the fluctuations in $\epsilon_{h,t}$ for different horizons h are perfectly correlated, and $\epsilon_{h,t}$ remains different from zero as $h \to \infty$, so that innovations in the error process bias expectations about dividend growth in the far future and not only in the near term.

These assumptions are all features of subjective forecasts of the future evolution of the state y_{t+h} in our model (if we identify our y_t with dividend growth). We have shown (in the right panel of Figure 6) that in our model, innovations in y_t cause subjective expectations of the future state to rise more than the RE forescast would, and the effect persists for several periods, though the bias caused by the innovation in any single period t eventually converges to zero. For each horizon h, (4.47) implies that the bias term is equal to $(1 - \rho^h)\hat{\mu}_t$; thus the biases for different forecast horizons are all perfectly correlated. Moreover, as the horizon is increased, the bias term becomes simply $\hat{\mu}_t$ for all large enough h; thus

the forecast errors predicted by the model are above all errors in long-term forecasts.

Our model also implies that there will be random fluctuations in forecast bias that are uncorrelated with any underlying fundamentals; these innovations are indicated by the $\tilde{\omega}_{t+1}$ shock in (4.46). The most important difference with the reduced-form specification of expectational bias proposed by Bordalo *et al.* is that in their model, there are arbitrary random variations in the "market expectations" that determine the value of the stock market; our model instead implies the existence of idiosyncratic random variation in the beliefs of an individual forecaster, but one might expect that these idiosyncratic variations should cancel out in their effects on the market price. It is possible that a satisfactory model of asset pricing will require us to suppose that some individual traders are large enough for their idiosyncratic beliefs to have a non-negligible effect on aggregate outcomes, as in the model of Gabaix *et al.* (2006). We leave the development of a complete model of asset prices for future work. But it seems likely that imperfect memory of the kind modeled here will be a necessary element in such a model.

8 Appendix

8.1 Reduction of the General Forecasting Problem to Estimation of μ

Consider the problem of choosing the vector of forecasts z_t each period so as to minimize (3.15). The elements of z_t must be chosen as a function of the DM's cognitive state at time t (after observing the external state y_t). As explained in the text, the DM's cognitive state at time t is assumed to consist of the value of the current external state y_t (observed with perfect precision), along with whatever additional information is reflected in the DM's period t memory state m_t. (In this section, it is not yet necessary to specify the nature of the vector m_t.)

If we use the notation $\mathrm{E}_t[\cdot]$ for the expectation of a random variable conditional on a complete description of the state at date t (including knowledge of the true value of μ), then

$$\mathrm{E}[(z_t - \mathrm{E}_t\tilde{z}_t)'W(\tilde{z}_t - \mathrm{E}_t\tilde{z}_t)] \;=\; 0,$$

since $\tilde{z}_t - \mathrm{E}_t\tilde{z}_t$ is a function of innovations in the external state subsequent to date t, that must be distributed independently of all of the determinants of both z_t and $\mathrm{E}_t\tilde{z}_t$. It follows that the term in (3.15) involving z_t can be equivalently expressed as[68]

$$\begin{aligned}\mathrm{E}[(z_t - \tilde{z}_t)'W(z_t - \tilde{z}_t)] &= \mathrm{E}[(z_t - \mathrm{E}_t\tilde{z}_t)'W(z_t - \mathrm{E}_t\tilde{z}_t)] \\ &\quad +\mathrm{E}[(\tilde{z}_t - \mathrm{E}_t\tilde{z}_t)'W(\tilde{z}_t - \mathrm{E}_t\tilde{z}_t)] \\ &\equiv L_{1t} + L_{2t}.\end{aligned}$$

Moreover, L_{2t} is independent of the decisions of the DM, and thus irrelevant to a determination of the optimal decision rule. The loss function (3.15) can thus equivalently be written as the discounted sum of the L_{1t} terms, which involve squared differences between z_t and $\mathrm{E}_t\tilde{z}_t$.

[68] Here we omit the factor β^t that multiplies this term in (3.15).

It further follows from the law of motion (3.14) that

$$\mathrm{E}_t \tilde{z}_t \;=\; \sum_{j=0}^{\infty} A_j[\mu + \rho^j(y_t - \mu)].$$

Since the precise value of y_t is presumed to be part of the cognitive state on the basis of which z_t can be chosen, one can write any decision rule in the form

$$z_t \;=\; \hat{z}_t \;+\; (\sum_{j=0}^{\infty} \rho^j A_j) \cdot y_t,$$

where $\hat{z}_t$ must be some function of the cognitive state at date t. In terms of this notation, the relevant part of the loss function (3.15) can then be written as

$$L_{1t} \;=\; \mathrm{E}[(\hat{z}_t - \mu a)' W (\hat{z}_t - \mu a)],$$

where we define $a \equiv \sum_{j=0}^{\infty}(1-\rho^j)A_j$ and make use of the fact that $\mathrm{E}_t[\mu] = \mu$.

The term L_{1t} that we wish to minimize can further be expressed as the expected value (integrating over all possible realizations of the cognitive state s_t in period t) of the quantity

$$\begin{aligned}
\tilde{L}_1(s_t) \;&\equiv\; \mathrm{E}[(\hat{z}_t - \mu a)' W (\hat{z}_t - \mu a)\,|s_t] \\
&=\; \mathrm{E}[\hat{z}_t\,|s_t]' W \mathrm{E}[\hat{z}_t\,|s_t] \;+\; \mathrm{E}[\breve{z}_t' W \breve{z}_t\,|s_t] \\
&\quad - 2a' W \mathrm{E}[\hat{z}_t\,|s_t] \cdot \mathrm{E}[\mu|s_t] \;+\; a' W a \cdot \mathrm{E}[\mu^2|s_t],
\end{aligned}$$

where we define $\breve{z}_t \equiv \hat{z}_t - \mathrm{E}[\hat{z}_t\,|s_t]$. (In expanding the right-hand side in this way, we use the fact that $\mathrm{E}[\breve{z}_t\,|s_t] = 0$, and that $\breve{z}_t$ must be independent of the deviation of μ from $\mathrm{E}[\mu|s_t]$, since the DM has no way to condition her action on μ except through the information about μ revealed by the cognitive state.) The expression $\tilde{L}_1(s_t)$ can then be separately minimized for each possible cognitive state s_t, by choosing a distribution for $\hat{z}_t$ conditional

on that state. We further note that the random component $\breve{z}_t$ of the action affects only the second term on the right-hand side, and so should be chosen to minimize that term; since W is positive definite, this is achieved by setting $\breve{z}_t = 0$ with certainty, so that $\hat{z}_t$ must be a deterministic function of s_t.

We can then simply write $\mathrm{E}[\hat{z}_t \,|s_t]$ as $\hat{z}_t$, and observe that

$$\tilde{L}_1(s_t) \;=\; (\hat{z}_t - a\mathrm{E}[\mu|s_t])'W(\hat{z}_t - a\mathrm{E}[\mu|s_t]) \;+\; a'Wa \cdot \mathrm{var}[\mu|s_t], \tag{H.58}$$

where the final term on the right-hand side is independent of the choice of $\hat{z}_t$. Thus in each cognitive state s_t, $\hat{z}_t$ must be chosen to minimize the first term on the right-hand side; since W is positive definite, this is achieved by setting $\hat{z}_t = a \cdot \hat{\mu}_t$, where $\hat{\mu}_t = \mathrm{E}[\mu|s_t]$.

Thus there is no loss of generality in restricting the DM to response rules of the form $\hat{z}_t = a \cdot \hat{\mu}_t$, where $\hat{\mu}_t$ is a scalar choice that depends on the cognitive state in period t, and that can be interpreted as the DM's estimate of μ given the cognitive state. Substituting this expression for $\hat{z}_t$ into (H.58), we have

$$\begin{aligned}
\tilde{L}_1(s_t) &= a'Wa \cdot \left\{(\hat{\mu}_t - \mathrm{E}[\mu|s_t])^2 \;+\; \mathrm{var}[\mu(s_t)]\right\} \\
&= a'Wa \cdot \mathrm{E}[(\hat{\mu}_t - \mu)^2 \,|s_t].
\end{aligned}$$

Then taking the unconditional expectation of this expression, we obtain

$$L_{1t} \;=\; \alpha \cdot MSE_t,$$

where $\alpha \;\equiv\; a'Wa \;>\; 0$ and MSE_t is defined as in the text.

Under any forecasting rule of the kind assumed here, then, the value of the loss function (3.15) will equal (3.17), plus an additional term

$$\sum_{t=0}^{\infty} \beta^t L_{2t}$$

that is independent of the DM's forecasting rule. Hence within this class of forecasting rules, the rule that minimizes (3.15) must be the one that minimizes (3.17); and since any other kind of forecasting rule can only lead to a higher value of (3.15), we can replace the problem of choosing a rule for determining z_t that minimizes (3.15) by the problem of choosing a rule for determining $\hat{\mu}_t$ that minimizes (3.17).

8.2 Bayesian Updating After the External State is Observed: A Kalman Filter

Let the elements of the memory state be partitioned as

$$m_t \;=\; \begin{bmatrix} \underline{m}_t \\ \\ \bar{m}_t \end{bmatrix}, \tag{H.59}$$

where the lower block consists of the elements of the reduced memory state

$$\bar{m}_t \;\equiv\; \mathrm{E}[x_t \,|m_t], \qquad \text{where } \; x_t \;\equiv\; \begin{bmatrix} \mu \\ \\ y_{t-1} \end{bmatrix},$$

while the upper block consists of the conditional expectations $\mathrm{E}[y_{t-j}\,|m_t]$ for $2 \leq j \leq t$. (This simply requires an appropriate ordering of the elements of m_t, using the notation for this vector introduced in the main text.)

We assume a posterior distribution of the form

$$x_t \,|m_t \;\sim\; N(\bar{m}_t,\, \Sigma_t)$$

conditional on the memory state m_t, where $\bar{m}_t$ is a 2-vector and Σ_t is a 2×2 symmetric, p.s.d. matrix. Under our assumption of linear-Gaussian dynamics for the memory state, the vector $\bar{m}_t$ will also be drawn from a multivariate Gaussian distribution. Since the prior

for the hidden state vector is specified to be

$$x_t \sim N(0, \Sigma_0), \qquad \Sigma_0 \equiv \begin{bmatrix} \Omega & \Omega \\ \Omega & \Omega + \sigma_y^2 \end{bmatrix}, \tag{H.60}$$

it follows that the unconditional distribution for the reduced memory state $\bar{m}_t$ must be of the form

$$\bar{m}_t \sim N(0, \Sigma_0 - \Sigma_t).$$

The complete set of variables (x_t, m_t) also have a multivariate Gaussian distribution. Moreover, since (by assumption) the expectation of x_t conditional on the realization of m_t depends only on the elements of $\bar{m}_t$, it follows that the entire distribution of x_t conditional on m_t depends only on $\bar{m}_t$, so that

$$x_t|m_t = x_t|\bar{m}_t.$$

Hence the joint distribution of the variables (x_t, m_t) can be factored as

$$p(x_t, \underline{m}_t, \bar{m}_t) = p(x_t, \bar{m}_t) \cdot p(\underline{m}_t\,|\bar{m}_t).$$

The DM then observes the external state y_t, which is assumed to depend on the hidden state vector x_t through an "observation equation" of the form

$$y_t = c'x_t + \epsilon_{yt}, \qquad \epsilon_{yt} \sim N(0, \sigma_\epsilon^2)$$

as a consequence of (3.14), where we further assume that ϵ_{yt} is distributed independently of both m_t and x_t. It follows that the variables (x_t, m_t, y_t) will have a joint distribution that

is multivariate Gaussian; and that this distribution can be factored as

$$\begin{aligned} p(x_t, m_t, y_t) &= p(x_t, m_t) \cdot p(y_t \,|x_t) \\ &= p(\underline{m}_t \,|\bar{m}_t) \cdot p(x_t, \bar{m}_t) \cdot p(y_t \,|x_t) \\ &= p(\underline{m}_t \,|\bar{m}_t) \cdot p(x_t, \bar{m}_t, y_t). \end{aligned}$$

From this it follows that

$$x_t \,|m_t, y_t \;=\; x_t \,|\bar{m}_t, y_t.$$

Thus both the expectation of x_t conditional on the cognitive state $s_t \equiv (m_t, y_t)$, and the variance-covariance matrix of the errors in the estimation of x_t based on the cognitive state, will depend only on the joint distribution of the variables $(x_t, \bar{m}_t, y_t)$. Moreover, the distribution for x_t conditional on the realizations of the elements of the cognitive state will be multivariate Gaussian,

$$x_t \,|\bar{m}_t, y_t \;\sim\; N(\bar{\mu}_t,\, \bar{\Sigma}_t), \tag{H.61}$$

where $\bar{\mu}_t$ is a linear function of $\bar{m}_t$ and y_t, while $\bar{\Sigma}_t$ is independent of the realizations of either $\bar{m}_t$ or y_t.

We can further decompose the vector of means $\bar{\mu}_t$ as

$$\begin{aligned} \bar{\mu}_t &= \mathrm{E}[x_t \,|\bar{m}_t, y_t] \\ &= \mathrm{E}[x_t \,|\bar{m}_t] \,+\, \{\mathrm{E}[x_t|\bar{m}_t, y_t] - \mathrm{E}[x_t|\bar{m}_t]\} \\ &= \bar{m}_t \,+\, \gamma_t \cdot (y_t \,-\, \mathrm{E}[y_t \,|\bar{m}_t]) \\ &= \bar{m}_t \,+\, \gamma_t \cdot (y_t \,-\, c'\mathrm{E}[x_t \,|\bar{m}_t]) \\ &= \bar{m}_t \,+\, \gamma_t \cdot (y_t \,-\, c'\bar{m}_t), \end{aligned}$$

where γ_t is the vector of *Kalman gains.* (The first element of this vector equation is then just equation (3.19) in the main text.)

The vector of Kalman gains must be chosen so that the estimation errors $x_t - \bar{\mu}_t$ are orthogonal to the surprise in the observation of the external state, $y_t - c'\bar{m}_t$. This requires that

$$\begin{aligned} 0 &= \mathrm{cov}(x_t - \bar{\mu}_t,\, y_t - c'\bar{m}_t) \\ &= \mathrm{cov}((x_t - \bar{m}_t) - \gamma_t(y_t - c'\bar{m}_t),\, y_t - c'\bar{m}_t) \\ &= \mathrm{var}[x_t - \bar{m}_t]c \,-\, \mathrm{var}[c'(x_t - \bar{m}_t) + \epsilon_{yt}] \cdot \gamma_t \\ &= \Sigma_t c \,-\, [c'\Sigma_t c + \sigma_\epsilon^2] \cdot \gamma_t. \end{aligned}$$

Hence

$$\gamma_t \;=\; \frac{\Sigma_t c}{c'\Sigma_t c + \sigma_\epsilon^2}. \tag{H.62}$$

The gain coefficient γ_{1t} in equation (3.19) is just the first element of this vector, $\gamma_{1t} \equiv e_1'\gamma_t$. This together with (H.62) yields the formula (3.21) given in the main text.

The variance-covariance matrix in the conditional distribution (H.61) will be given by

$$\begin{aligned} \bar{\Sigma}_t &= \mathrm{var}[x_t - \bar{\mu}_t] \;=\; \mathrm{var}[(x_t - \bar{m}_t) - \gamma_t(y_t - c'\bar{m}_t)] \\ &= \mathrm{var}[(I - \gamma_t c')(x_t - \bar{m}_t) \,-\, \gamma_t\epsilon_{yt}] \\ &= (I - \gamma_t c')\Sigma_t(I - \gamma_t c')' \,+\, \sigma_\epsilon^2\gamma_t\gamma_t' \\ &= \Sigma_t \,-\, 2[c'\Sigma_t c + \sigma_\epsilon^2]\gamma_t\gamma_t' \,+\, [c'\Sigma_t c]\gamma_t\gamma_t' \,+\, \sigma_\epsilon^2\gamma_t\gamma_t' \\ &= \Sigma_t \,-\, [c'\Sigma_t c + \sigma_\epsilon^2]\gamma_t\gamma_t'. \end{aligned}$$

The remaining uncertainty about the value of μ given the cognitive state, $\hat{\sigma}_t^2$, is then equal to $\bar{\Sigma}_{11,t}$, so that

$$\hat{\sigma}_t^2 \;=\; e_1'\bar{\Sigma}_t e_1 \;=\; e_1'\Sigma_t e_1 \,-\, (c'\Sigma_t c + \sigma_\epsilon^2)(\gamma_{1t})^2,$$

which is just expression (3.20) in the main text.

Substituting expression (H.60) for Σ_0 into this solution, we obtain

$$\begin{aligned}\hat{\sigma}_0^2 &= \Omega - \left(\Omega + \sigma_y^2\right) \cdot \left[\frac{\Omega}{\Omega + \sigma_y^2}\right]^2 \\ &= \frac{\Omega \sigma_y^2}{\Omega + \sigma_y^2},\end{aligned}$$

which is the formula given in (3.22). It remains to be shown that this is an upper bound for $\hat{\sigma}_t^2$. To show this, we observe that

$$\begin{aligned}\hat{\sigma}_t^2 &= \min_{\beta,\gamma_1} \text{var}[\mu - \beta' \bar{m}_t - \gamma_1 y_t] \\ &\leq \min_{\gamma_1} \text{var}[\mu - \gamma_1 y_t] \\ &\leq \text{var}[\mu - (\Omega/(\Omega + \sigma_y^2)) \cdot y_t] \\ &= \text{var}[(\sigma_y^2/(\Omega + \sigma_y^2))\mu - (\Omega/(\Omega + \sigma_y^2))(y_t - \mu)] \\ &= \left(\frac{\sigma_y^2}{\Omega + \sigma_y^2}\right)^2 \text{var}[\mu] + \left(\frac{\Omega}{\Omega + \sigma_y^2}\right)^2 \text{var}[y_t|\mu] \\ &= \left(\frac{\sigma_y^2}{\Omega + \sigma_y^2}\right)^2 \Omega + \left(\frac{\Omega}{\Omega + \sigma_y^2}\right)^2 \sigma_y^2 \\ &= \frac{\Omega \sigma_y^2}{\Omega + \sigma_y^2} = \sigma_0^2.\end{aligned}$$

This establishes the upper bound (3.22) stated in the main text.

8.3 Demonstration that an Optimal Memory Structure Records Information Only about the Reduced Cognitive State

Let (2.2) be written in the partitioned form

$$\begin{bmatrix} \underline{m}_{t+1} \\ \\ \bar{m}_{t+1} \end{bmatrix} = \begin{bmatrix} \Lambda_{a,t} & \Lambda_{b,t} \\ \\ \Lambda_{c,t} & \Lambda_{d,t} \end{bmatrix} \begin{bmatrix} \underline{s}_t \\ \\ \bar{s}_t \end{bmatrix} + \begin{bmatrix} \underline{\omega}_{t+1} \\ \\ \bar{\omega}_{t+1} \end{bmatrix}. \tag{H.63}$$

Here m_{t+1} is again partitioned as in (H.59). The lower block of s_t consists of the elements of the reduced cognitive state

$$\bar{s}_t \equiv \begin{bmatrix} \hat{\mu}_t \\ \\ y_t \end{bmatrix},$$

both elements of which are linear functions of s_t, as a consequence of equation (3.19). We choose a representation for the vector s_t such that the lower block consists of the elements of $\bar{s}_t$, the elements of $\underline{s}_t$ are all uncorrelated with the elements of $\bar{s}_t$, and the elements of the vectors $\bar{s}_t$ and $\underline{s}_t$ together span the same linear space of random variables as the elements of s_t. (We can necessarily write any memory structure of the form (2.2) in this way; it amounts simply to a choice of the basis vectors in terms of which the vectors m_{t+1} and s_t are each decomposed.)

Let us suppose furthermore that a representation for m_{t+1} is chosen consistent with the normalization $\mathrm{E}[\bar{s}_t \,|m_{t+1}] = \bar{m}_{t+1}$. This holds if and only if both elements of the vector $\bar{s}_t - \bar{m}_{t+1}$ are uncorrelated with each of the elements of m_{t+1}. These consistency conditions can be reduced to two requirements: (i) the requirement that

$$\mathrm{var}[\Lambda_{c,t}\underline{s}_t + \bar{\omega}_{t+1}] = (I - \Lambda_{d,t})X_t\Lambda'_{d,t}, \tag{H.64}$$

where the matrix $X_t \equiv \mathrm{var}[\bar{s}_t]$ is independent of the memory structure chosen for period t; and (ii) the requirement that $\bar{s}_t - \bar{m}_{t+1}$ be uncorrelated with all elements of $\underline{m}_{t+1}$. (Note that $\bar{s}_t - \bar{m}_{t+1}$ is uncorrelated with $\bar{m}_{t+1}$ if and only if (H.64) holds.)

Forecast accuracy depends only on the matrices $\{\Lambda_{d,t}\}$

Suppose that in any period t, we take the memory structure in periods $\tau < t$ as given. This means that the DM's uncertainty about x_t given the memory state m_t (specified by the posterior variance-covariance matrix Σ_t) will be given. (If $t = 0$, Σ_0 is simply given by the

prior.) Hence the value of $\hat{\mu}_t$ as a function of $\bar{m}_t$ and y_t will be given, and consequently the value of MSE_t will be given, following the discussion in the main text (and the previous section of this appendix). The elements of the matrix X_t will similarly be given.

We next consider how $\Lambda_{d,t}$ must be chosen, in order for it to be possible to choose matrices $\Lambda_{c,t}$ and $\mathrm{var}[\bar{\omega}_{t+1}]$ such that (H.64) is satisfied. Equation (H.64) requires that $(I-\Lambda_{d,t})X_t\Lambda'_{d,t}$, be a symmetric matrix; this will hold if and only if the simpler requirement is satisfied that $\Lambda_{d,t}X_t = X_t\Lambda'_{d,t}$ be a symmetric matrix. In addition, it is necessary that $(I-\Lambda_{d,t})X_t\Lambda'_{d,t}$ be a p.s.d. matrix. The set of matrices $\Lambda_{d,t}$ with these properties is a non-empty set ($\Lambda_{d,t} = 0$ is a trivial example), and depends only on the matrix X_t. Let this set of matrices be denoted $\mathcal{L}(X_t)$.

Now let $\Lambda_{d,t}$ be any matrix that belongs to $\mathcal{L}(X_t)$. Then it is possible to choose the matrices $\Lambda_{c,t}$ and $\mathrm{var}[\bar{\omega}_{t+1}]$ so that (H.64) is satisfied; and given any such choice of these two matrices, it is further possible to choose the specification of the equation for $\underline{m}_{t+1}$ so that all elements of $\underline{m}_{t+1}$ are uncorrelated with the elements of $\bar{s}_t - \bar{m}_{t+1}$. Given any such specifications, both conditions (i) and (ii) above will be satisfied. Thus the matrix $\Lambda_{d,t}$ is admissible as part of the specification of a memory structure; and any possible memory structure consistent with the matrix $\Lambda_{d,t}$ will be one of those with the properties just assumed.

Given a matrix $\Lambda_{d,t}$ of this sort, we next observe that the equations determining $\bar{m}_{t+1}$ can be written in the form

$$\bar{m}_{t+1} \;=\; \Lambda_{d,t}\bar{s}_t \;+\; \nu_{t+1},$$

where $\nu_{t+1} \sim N(0,\, \Lambda_{d,t}X_t)$ is distributed independently of $\bar{s}_t$. Thus the joint distribution of $(\bar{s}_t, \bar{m}_{t+1})$ will be a multivariate Gaussian distribution, the parameters of which are completely determined by X_t and $\Lambda_{d,t}$. It then follows that the conditional distribution $\bar{s}_t|\bar{m}_{t+1}$ will be a bivariate Gaussian distribution, with a mean $\bar{m}_{t+1}$ and a variance independent of the realization of $\bar{m}_{t+1}$, which also depends only on X_t and $\Lambda_{d,t}$. Moreover, since the elements of $\underline{m}_{t+1}$ are all Gaussian random variables distributed independently of $\bar{s}_t - \bar{m}_{t+1}$,

knowledge of $\underline{m}_{t+1}$ cannot further improve one's estimate of $\bar{s}_t$, and so the conditional distribution $\bar{s}_t|m_{t+1} = \bar{s}_t|\bar{m}_{t+1}$. Finally, since we can write

$$x_{t+1} \;=\; \bar{s}_t \;+\; \begin{bmatrix} \\ u_t \\ \\ 0 \end{bmatrix},$$

where $u_t \;\sim\; N(0,\, \hat{\sigma}_t^2)$ must be uncorrelated with any of the elements of s_t (and hence uncorrelated with any of the elements of m_{t+1}), we must further have

$$x_{t+1}|m_{t+1} \;\sim\; N(\bar{m}_{t+1},\, \Sigma_{t+1})$$

where

$$\Sigma_{t+1} \;=\; \mathrm{var}[\bar{s}_t\,|\bar{m}_{t+1}] \;+\; \hat{\sigma}_t^2\, e_1 e_1'.$$

Since $\hat{\sigma}_t^2$ also depends only on Σ_t (see equation (3.20)), it follows that the elements of Σ_{t+1} depend only on Σ_t and $\Lambda_{d,t}$.

This argument can then be used recursively (starting from period $t = 0$) to show that given the initial uncertainty matrix Σ_0 implied by the prior (H.60), we can completely determine the entire sequence of matrices $\{\Sigma_t\}$, given a sequence of matrices $\{\Lambda_{d,t}\}$ for all $t \geq 0$ with the property that for each t, $\Lambda_{d,t} \in \mathcal{L}(X_t)$, where X_t is the matrix implied by Σ_t. Moreover, given such a sequence of matrices $\{\Lambda_{d,t}\}$, the value of MSE_t for each period t will be uniquely determined as well. Hence the terms in the loss function (3.18) that depend on the accuracy of forecasts that are possible using a given memory structure will depend only on the sequence of matrices $\{\Lambda_{d,t}\}$. (These matrices must be chosen to satisfy a set of consistency conditions, stated above, but these conditions can also be expressed purely in terms of the sequence of matrices $\{\Lambda_{d,t}\}$.) Thus the other elements of the specification (H.63) of the memory structure matter only to the extent that they have consequences for the information cost terms in (3.18).

Mutual information: a useful lemma

Information costs in period t are assumed to be an increasing function of $I_t = I(M; S)$, the Shannon mutual information between random variables M (the realizations of which are denoted m_{t+1}) and S (the realizations of which are denoted s_t).[69] Each of the random vectors M and S can further be partitioned as $M = (\underline{M}, \bar{M})$, $S = (\underline{S}, \bar{S})$.

Now for any random variables $X_1, X_2, \ldots$, let $H(X_1, X_2, \ldots, X_k)$ be the entropy of the joint distribution for variables $(X_1, X_2, \ldots, X_k)$, and $H(X_1, \ldots, X_k \,|X_{k+1}, \ldots X_{k+m})$ be the entropy of the joint distribution of the variables $(X_1, \ldots, X_k)$ conditional on the values of the variables $(X_{k+1}, \ldots X_{k+m})$. The chain rule for entropy implies that

$$H(X_1, X_2, \ldots, X_k) \;=\; H(X_1) \;+\; H(X_2 \,|X_1) \;+\; \ldots \;+\; H(X_k \,|X_1, \ldots, X_{k-1}).$$

We can then define the mutual information between the variables $(X_1, \ldots, X_k)$ and the variables $(X_{k+1}, \ldots X_{k+m})$ as

$$I(X_1, \ldots, X_k;\, X_{k+1}, \ldots, X_{k+m}) \;\equiv\; H(X_1, \ldots, X_k) \;-\; H(X_1, \ldots, X_k \,|X_{k+1}, \ldots X_{k+m}).$$

(The information about the first set of variables that is revealed by learning the values of the second set of variables is measured by the average amount by which the entropy of the conditional distribution is smaller than the entropy of the unconditional distribution of the first set of variables.) Similarly, we can define the mutual information between the first set of variables and the second set of variables, conditioning on the values of some third set of variables as

$$I(X_1, \ldots, X_k;\, X_{k+1}, \ldots, X_{k+m} \,|X_{k+m+1}, \ldots, X_{k+m+n})$$

[69] Here we adopt the notation used in Cover (2006), with different symbols for the random variables M and S and their realizations. This is to make it clear that I_t is not a function of the values taken by m_{t+1} and s_t along a particular history, but instead a function of the complete joint distribution of the two random variables; I_t is itself not a random variable, but a single number for each date t.

$$\equiv\ H(X_1, X_2, \ldots, X_k\,|X_{k+m+1}, \ldots, X_{k+m+n})\ -\ H(X_1, \ldots, X_k\,|X_{k+1}, \ldots, X_{k+m+n}).$$

Thus for any set of four random variables $\underline{M}, \bar{M}, \underline{S}, \bar{S}$, we must have

$$\begin{aligned}
I(\underline{S}, \bar{S};\, \underline{M}, \bar{M}) &= H(\underline{S}, \bar{S})\ -\ H(\underline{S}, \bar{S}\,|\underline{M}, \bar{M}) \\
&= [H(\bar{S}) + H(\underline{S}\,|\bar{S})]\ -\ [H(\bar{S}\,|\underline{M}, \bar{M}) + H(\underline{S}\,|\bar{S}, \underline{M}, \bar{M})] \\
&= [H(\bar{S}) + H(\underline{S}\,|\bar{S})]\ -\ [H(\bar{S}, \underline{M}, \bar{M}) - H(\underline{M}\,|\bar{M}) - H(\bar{M})]\ -\ H(\underline{S}\,|\bar{S}, \underline{M}, \bar{M}) \\
&= [H(\bar{S}) + H(\underline{S}\,|\bar{S})]\ -\ [(H(\bar{M}) + H(\bar{S}\,|\bar{M}) + H(\underline{M}\,|\bar{M}, \bar{S})) - H(\underline{M}\,|\bar{M}) - H(\bar{M})] \\
&\qquad\qquad - H(\underline{S}\,|\bar{S}, \underline{M}, \bar{M}) \\
&= [H(\bar{S}) + H(\underline{S}\,|\bar{S})]\ -\ [H(\bar{S}\,|\bar{M}) + H(\underline{M}\,|\bar{M}, \bar{S}) - H(\underline{M}\,|\bar{M})]\ -\ H(\underline{S}\,|\bar{S}, \underline{M}, \bar{M}) \\
&= [H(\bar{S}) - H(\bar{S}\,|\bar{M})]\ +\ [H(\underline{S}\,|\bar{S}) - H(\underline{S}\,|\bar{S}, \underline{M}, \bar{M})]\ +\ [H(\underline{M}\,|\bar{M}) - H(\underline{M}\,|\bar{M}, \bar{S})] \\
&= I(\bar{S};\, \bar{M})\ +\ I(\underline{S};\, \underline{M}, \bar{M}\,|\bar{S})\ +\ I(\underline{M};\, \bar{S}\,|\bar{M}).
\end{aligned}$$

Then, since mutual information is necessarily non-negative, we can establish the lower bound

$$I_t\ =\ I(\underline{S}, \bar{S};\, \underline{M}, \bar{M})\ \geq\ I(\bar{S};\, \bar{M}). \tag{H.65}$$

Furthermore, this lower bound is achieved if and only if

$$I(\underline{S};\, \underline{M}, \bar{M}\,|\bar{S})\ =\ I(\underline{M};\, \bar{S}\,|\bar{M})\ =\ 0.$$

For any three random variables X, Y, Z, the conditional mutual information $I(X;\, Y\,|Z) = 0$ if and only if the variables X and Y are distributed independently one another, conditional on the value of Z. Hence the lower bound (H.65) is achieved if and only if (a) conditional on the value of $\bar{m}_{t+1}$, the variables $\bar{s}_t$ and $\underline{m}_{t+1}$ are independent of one another; and (b) conditional on the value of $\bar{s}_t$, the variables $\underline{s}_t$ and m_{t+1} are independent of one another.

Optimality of Setting $\Lambda_{a,t} = \Lambda_{b,t} = \Lambda_{c,t} = 0$

We return now to the consideration of possible memory structures. Let the sequence of matrices $\{\Lambda_{d,t}\}$ be chosen to satisfy the consistency conditions discussed above, and for a given such sequence, consider an optimal choice of the remaining elements of the specification (H.63), from among those specifications that are consistent with the sequence $\{\Lambda_{d,t}\}$ (that is, that will satisfy both conditions (i) and (ii) stated above).

We have shown above that the sequence of values $\{MSE_t\}$ is completely determined by the specification of $\{\Lambda_{d,t}\}$. Hence other aspects of the specification of the memory structure can matter only to the extent that they affect the sequence of values $\{I_t\}$. Moreover, we have shown that the joint distribution of $(\bar{s}_t, \bar{m}_{t+1})$ each period is completely determined by X_t and $\Lambda_{d,t}$, which means that the lower bound for I_t given in (H.65) is completely determined by the choice of $\{\Lambda_{d,\tau}\}$ for $\tau \leq t$. It thus remains only to consider whether this lower bound can be achieved, and under what conditions.

We first observe that the lower bound is achievable. For any sequence of matrices $\{\Lambda_{d,t}\}$ satisfying the specified conditions, a memory structure specification with $\Lambda_{a,t} = \Lambda_{b,t} = \Lambda_{c,t} = 0$, together with a stipulation that $\underline{\omega}_{t+1}$ be distributed independently of $\bar{\omega}_{t+1}$ and that $\text{var}[\bar{\omega}_{t+1}] = \Lambda_{d,t} X_t$, will satisfy both conditions (i) and (ii) stated in the introduction to this appendix, and thus this represents a feasible memory structure. One can also show that such a specification satisfies both of conditions (a) and (b) stated at the end of section C.2, so that the lower bound (H.65) is achieved in each period. Thus such a specification achieves the lowest possible value for the combined objective function (3.18), and will be optimal, given our choice of the sequence $\{\Lambda_{d,t}\}$.

Not only will this specification be sufficient for achieving the lowest possible value of (3.18), but it will be essentially necessary. We have shown above that achieving the lower bound for I_t in period t requires that conditional on the value of $\bar{s}_t$, the variables $\underline{s}_t$ and m_{t+1} are independent of one another. This means that the values of the variables in the vector $\underline{s}_t$ cannot help at all in predicting any elements of m_{t+1}, once one is already using

the reduced cognitive state $\bar{s}_t$ to forecast the next period's memory state; thus one must be able to write law of motion (H.63) for the memory state with $\Lambda_{a,t} = \Lambda_{c,t} = 0$.[70] Thus it is necessarily the case that the elements of m_{t+1} convey information only about the reduced cognitive state $\bar{s}_t$, and not about any other aspects of the cognitive state s_t.

In addition, we have shown above that achieving the lower bound for I_t in period t requires that conditional on the value of $\bar{m}_{t+1}$, the variables $\bar{s}_t$ and $\underline{m}_{t+1}$ are independent of one another. Thus all of the information about $\bar{s}_t$ that is contained in the memory state m_{t+1} is contained in the elements $\bar{m}_{t+1}$. This means either that $\Lambda_{b,t} = 0$ as well, or, to the extent that some element of $\underline{m}_{t+1}$ corresponds to a row of $\Lambda_{b,t}$ with non-zero elements, that element of $\underline{m}_{t+1}$ must be a linear combination of the elements of $\bar{m}_{t+1}$, so that conditioning upon its value conveys no new information about $\bar{s}_t$. Thus any specification of the memory structure in which $\Lambda_{b,t} \neq 0$ in any period represents a redundant representation of the contents of memory available in period $t+1$; we can equivalently describe the contents of memory by eliminating all such rows from m_{t+1}.

Thus there is no loss of generality in assuming that the lower bound is achieved by specifying $\Lambda_{a,t} = \Lambda_{b,t} = \Lambda_{c,t} = 0$ in each period. Finally, satisfaction of consistency condition (ii) in this case requires that the elements of $\underline{\omega}_{t+1}$ be distributed independently of the elements of $\bar{\omega}_{t+1}$. We might still allow $\mathrm{var}[\underline{\omega}_{t+1}]$ to be non-zero; this would mean that $\underline{m}_{t+1}$ contains elements that fluctuate randomly, but are completely uncorrelated with the previous period's cognitive state s_t. Such an information structure is equally optimal, in the sense that (3.18) is made no larger by the existence of such components of the memory state, given our assumption that only mutual information is costly. But the additional components $\underline{m}_{t+1}$ of the memory structure will have no consequences for cognitive processing, and our inclusion of them as part of the representation of the memory state violates our

[70] It might be possible to satisfy the condition required for the lower bound with non-zero elements in one of these matrices; but this will occur only because of collinearity in the fluctuations in the elements of the vector $\underline{s}_t$, so that it is possible to have a law of motion in which $\underline{s}_t$ has no effect on m_{t+1}, despite non-zero matrices $\Lambda_{a,t}$ and $\Lambda_{c,t}$. In such a case, the representation of the cognitive state by the vector s_t would involve redundancy; and in any event, there would be no loss of generality in setting $\Lambda_{a,t} = \Lambda_{c,t} = 0$, since the implied fluctuations in the memory state would be the same.

assumption in the text that we label memory states by their implied posteriors for the values of μ and the past realizations of the external state; using labels $(\underline{m}_{t+1}, \bar{m}_{t+1})$ in which $\underline{m}_{t+1}$ is non-null will mean having separate labels for memory states that imply the same posterior (since the value of $\underline{m}_{t+1}$ would be completely uninformative about either μ or any past external states).

Hence in the case of any optimal memory structure, the memory state can be described more compactly by identifying it with the reduced memory state $\bar{m}_{t+1}$, which evolves according to

$$\bar{m}_{t+1} \;=\; \bar{\Lambda}_t \bar{s}_t \;+\; \bar{\omega}_{t+1}, \tag{H.66}$$

where $\bar{\Lambda}_t$ is the matrix called $\Lambda_{d,t}$ in (H.63). (This corresponds to equation (3.25) in the main text.) We need only consider (at most) a two-dimensional memory state, and the optimal memory state conveys information only about the reduced cognitive state $\bar{s}_t$, not about any other aspects of the cognitive state s_t.

An alternative representation for the reduced cognitive state

We have shown in the main text (equation (3.28)) that the variance matrix of the reduced cognitive state $\bar{s}_t$ can be written as a function of the single parameter $\hat{\sigma}_t^2$:

$$X_t \;=\; X(\hat{\sigma}_t^2) \;\equiv\; \begin{bmatrix} \Omega - \hat{\sigma}_t^2 & \Omega \\ \Omega & \Omega + \sigma_y^2 \end{bmatrix}.$$

There is another way of writing this function that will be useful below.

We can orthogonalize the reduced cognitive state using the transformation $\bar{s}_t = \Gamma \check{s}_t$, where

$$\Gamma \;\equiv\; \begin{bmatrix} 1 & \frac{\Omega}{\Omega+\sigma_y^2} \\ 0 & 1 \end{bmatrix}. \tag{H.67}$$

The elements of the orthogonalized cognitive state have the interpretation

$$\check{s}_t \equiv \begin{bmatrix} \hat{\mu}_t - \mathrm{E}[\mu|y_t] \\ \\ y_t \end{bmatrix},$$

from which it is obvious that the first element must be uncorrelated with the second.

The variance matrix of $\check{s}_t$ is therefore diagonal:

$$\mathrm{var}[\check{s}_t] = \check{X}(\hat{\sigma}_t^2) \equiv \begin{bmatrix} \hat{\sigma}_0^2 - \hat{\sigma}_t^2 & 0 \\ \\ 0 & \Omega + \sigma_y^2 \end{bmatrix}. \tag{H.68}$$

We can then alternatively write

$$X(\hat{\sigma}_t^2) = \Gamma \check{X}(\hat{\sigma}_t^2)\Gamma'. \tag{H.69}$$

8.4 The Law of Motion for the Memory State and the Information Content of Memory

We now consider how the parameterization of the law of motion (H.66) for the memory state determines the degree of uncertainty about the external state vector that will exist when beliefs are conditioned on the memory state, and how the same parameters determine the mutual information between the memory state and the prior cognitive state, and hence the size of the information cost term $c(I_t)$.

We begin by recapitulating the conditions that the sequence of matrices $\{\bar{\Lambda}_t\}$ and $\{\Sigma_{\bar{\omega},t+1}\}$ must satisfy, in order for (H.66) to represent a memory structure consistent with the normalization according to which $\mathrm{E}[x_{t+1}\,|\bar{m}_{t+1}] = \bar{m}_{t+1}$. Condition (H.64) will be satisfied if and only if

$$\Sigma_{\bar{\omega},t+1} = (I - \bar{\Lambda}_t)X_t\bar{\Lambda}_t'. \tag{H.70}$$

In order for there to be a symmetric, p.s.d. matrix $\Sigma_{\bar{\omega},t+1}$ that satisfies (H.70), it must be the case that $\bar{\Lambda}_t \in \mathcal{L}(X_t)$. As explained above, this means that $\bar{\Lambda}_t X_t = X_t \bar{\Lambda}_t'$ must be a symmetric matrix, and in addition that $(I - \bar{\Lambda}_t) X_t \bar{\Lambda}_t'$ is p.s.d. Note that since

$$X_t \bar{\Lambda}_t' \;=\; (I - \bar{\Lambda}_t) X_t \bar{\Lambda}_t' \;+\; \bar{\Lambda}_t X_t \bar{\Lambda}_t',$$

and X_t is necessarily a p.s.d. matrix, it follows from the assumption that $(I - \bar{\Lambda}_t) X_t \bar{\Lambda}_t'$ is p.s.d. that $\bar{\Lambda}_t X_t = X_t \bar{\Lambda}_t'$ will also be a p.s.d. matrix; but this latter condition is weaker than the one assumed in our definition of the set $\mathcal{L}(X_t)$. This constitutes the complete set of conditions that must be satisfied for (H.66) to represent a memory structure consistent with our proposed normalization of the vector m_{t+1}.

We can further specialize these conditions in the case that $\bar{\Lambda}_t$ is a singular matrix. (Here we assume that X_t is of full rank.) If $\bar{\Lambda}_t$ is of rank one (or less), it can be written in the form $\bar{\Lambda}_t = u_t v_t'$, where we are furthermore free to normalize the vector v_t' so that $v_t' X_t v_t = 1$. Then the condition that $\bar{\Lambda}_t X_t = X_t \bar{\Lambda}_t'$ will hold only if $u_t (v_t' X_t) = (X_t v_t) u_t'$. This means that u_t must be collinear with $X_t v_t$, so that we must be able to write $u_t = \lambda_t X_t v_t$, for some scalar λ_t. Thus in the singular case, we must be able to write

$$\bar{\Lambda}_t \;=\; \lambda_t X_t v_t v_t', \tag{H.71}$$

where λ_t is a scalar and v_t is a vector such that $v_t' X_t v_t = 1$. Then

$$(I - \bar{\Lambda}_t) X_t \bar{\Lambda}_t' \;=\; \lambda_t (1 - \lambda_t)(X_t v_t)(X_t v_t)'$$

will be a p.s.d. matrix if and only if in addition $0 \leq \lambda_t \leq 1$. Thus a singular matrix $\bar{\Lambda}_t$ is an element of $\mathcal{L}(X_t)$ if and only if it is of the form (H.71) with $0 \leq \lambda_t \leq 1$ and v_t a vector such that $v_t' X_t v_t = 1$.

Consistency with the proposed normalization of m_{t+1} then further requires that

$$\Sigma_{\bar{\omega},t+1} \;=\; \lambda_t(1-\lambda_t)X_t v_t v_t' X_t. \tag{H.72}$$

This implies that $\Sigma_{\bar{\omega},t+1}$ is a singular matrix; the random vector $\bar{\omega}_{t+1}$ can be written as $\bar{\omega}_{t+1} \;=\; X_t v_t \cdot \tilde{\omega}_{t+1}$, where $\tilde{\omega}_{t+1}$ is a scalar random variable, with distribution $N(0,\, \lambda_t(1-\lambda_t)$. It follows that in such a case, the memory state can be given a one-dimensional representation, writing $\bar{m}_{t+1} \;=\; X_t v_t \cdot \tilde{m}_{t+1}$, where the scalar memory state $\tilde{m}_{t+1}$ has a law of motion

$$\tilde{m}_{t+1} \;=\; \lambda_t v_t' \bar{s}_t \;+\; \tilde{\omega}_{t+1}, \qquad \tilde{\omega}_{t+1} \;\sim\; N(0,\, \lambda_t(1-\lambda_t)). \tag{H.73}$$

In the case that $X_t = X_0$ (the only case in which it is possible for $X_t = X(\hat{\sigma}_t^2)$ to be singular), we have defined $\mathcal{L}(X_0)$ to include only matrices of the special form (3.29) with $0 \leq \lambda_t \leq 1$. In this case, $\bar{\Lambda}_t$ is necessarily of the form (H.71), with the vector v_t given by (3.38). Hence our comments above about the case in which $\bar{\Lambda}_t$ is singular apply also in the case in which X_t is singular, except that in this latter case we have the further restriction that v_t must be given by (3.38). In this special case, (H.72) reduces to

$$\Sigma_{\bar{\omega},t+1} \;=\; \lambda_t(1-\lambda_t)[\Omega + \sigma_y^2]\, ww'.$$

The degree of uncertainty implied by a given memory structure

We turn now to the question of how the posterior uncertainty Σ_{t+1} in the following period is determined by the law of motion for the memory state $\bar{m}_{t+1}$ that can be accessed at that time. Note that the variance of the marginal distribution for x_{t+1} can be decomposed as

$$\mathrm{var}[x_{t+1}] \;=\; \mathrm{E}[\mathrm{var}[x_{t+1}\,|m_{t+1}]] \;+\; \mathrm{var}[\mathrm{E}[x_{t+1}\,|m_{t+1}]],$$

where in the first term on the right-hand side, the variance refers to the distribution of values for x_{t+1} conditional on the realization of m_{t+1}, and the expectation is over realizations of m_{t+1}, while in the second term the variance refers to the distribution of values for m_{t+1}, and the expectation is over values of x_{t+1} conditional on the realization of m_{t+1}. Since the marginal distribution for x_{t+1} is the same for all t, and coincides with the prior distribution for x_0 specified in (H.60), the left-hand side must equal the matrix Σ_0 defined there. Hence the variance decomposition can be written as

$$\Sigma_0 \;=\; \Sigma_{t+1} \;+\; \mathrm{var}[\bar{m}_{t+1}],$$

which implies that in any period,

$$\Sigma_{t+1} \;=\; \Sigma_0 \;-\; \mathrm{var}[\bar{m}_{t+1}].$$

Thus in order to understand how the choice of $\bar{\Lambda}_t$ determines Σ_{t+1}, it suffices that we determine the implications for the degree of variation in $\bar{m}_{t+1}$.

A law of motion of the form (H.66) implies that

$$\begin{aligned}
\mathrm{var}[\bar{m}_{t+1}] \;&=\; \bar{\Lambda}_t X_t \bar{\Lambda}_t' \;+\; \Sigma_{\bar{\omega},t+1} \\
&=\; \bar{\Lambda}_t X_t \bar{\Lambda}_t' \;+\; (I - \bar{\Lambda}_t) X_t \bar{\Lambda}_t' \\
&=\; X_t \bar{\Lambda}_t',
\end{aligned}$$

where the second line uses (H.70). Hence we obtain the prediction that

$$\Sigma_{t+1} \;=\; \Sigma_0 \;-\; X_t \bar{\Lambda}_t'. \tag{H.74}$$

Note that for any $\bar{\Lambda}_t \in \mathcal{L}(X_t)$, this must be a symmetric, p.s.d. matrix.

Hence for any value of $\hat{\sigma}_t^2$ satisfying $0 \leq \hat{\sigma}_t^2 \leq \hat{\sigma}_0^2$ and any transition matrix $\bar{\Lambda}_t \in$

$\mathcal{L}(X(\hat{\sigma}_t^2))$, we can substitute $X_t = X(\hat{\sigma}_t^2)$ and the value of Σ_{t+1} given by (H.74) into (3.20) to obtain a solution for $\hat{\sigma}_{t+1}^2$ as a function of $\hat{\sigma}_t^2$ and $\bar{\Lambda}_t$. This defines the function $f(\hat{\sigma}_t^2, \bar{\Lambda}_t)$ referred to in the main text. We can then define $\mathcal{L}^{seq}$ as the set of sequences of transition matrices $\{\bar{\Lambda}_t\}$ for all $t \geq 0$ such that

$$\bar{\Lambda}_0 \in \mathcal{L}(X_0), \qquad \bar{\Lambda}_1 \in \mathcal{L}(X(f(\hat{\sigma}_0^2, \bar{\Lambda}_0))), \qquad \bar{\Lambda}_2 \in \mathcal{L}(X(f(f(\hat{\sigma}_0^2, \bar{\Lambda}_0), \bar{\Lambda}_1))),$$

and so on.

Then given any sequence of transition matrices $\{\bar{\Lambda}_t\} \in \mathcal{L}^{seq}$, there will be uniquely defined sequences $\{\hat{\sigma}_t^2, X_t\}$ for all $t \geq 0$. Equation (H.74), together with (H.60), can then be used to uniquely define the implied sequence of matrices $\{\Sigma_t\}$ for all $t \geq 0$. These matrices can in turn be used in (3.21) to define the Kalman gain γ_{1t} for each $t \geq 0$. Thus for any sequence of transition matrices $\{\bar{\Lambda}_t\} \in \mathcal{L}^{seq}$, there will be uniquely determined sequences $\{\Sigma_t, \gamma_{1t}, \hat{\sigma}_t^2, X_t\}$, as stated in the text. These in turn will imply a uniquely determined sequence of losses $\{MSE_t\}$ from forecast inaccuracy, using (3.23).

The mutual information implied by a given memory structure

Finally, we compute the mutual information I_t in the case that the memory state consists only of a reduced memory state $\bar{m}_{t+1}$, with law of motion (H.66). We first review the definition of mutual information in the case of continuously distributed random variables.

Let X and Y be two random variables, each parameterized using a finite system of coordinates (so that realizations x and y are each represented by finite-dimensional vectors), and suppose that at least Y has a continuous distribution, with a density function $p(y|x)$ such that $p(y|x) > 0$ for all y in the support of Y and all x in the support of X. Suppose also that the marginal distribution for Y can be characterized by a density function $p(y) = \mathrm{E}[p(Y|x)]$, where the expectation is over possible realizations of x, and $p(y) > 0$ for all y in the support of Y. Then we can measure the degree to which knowing the

realization of x changes the distribution that one can expect y to be drawn from by the Kullback-Liebler divergence (or relative entropy) of the conditional distribution $p(y|x)$ relative to the marginal distribution $p(y)$, defined as

$$D_{KL}(p(\cdot|x)||p(\cdot)) \equiv \mathrm{E}\left[\log \frac{p(y|x)}{p(y)}\right] \geq 0, \tag{H.75}$$

where the expectation is over possible realizations of y, and this quantity is a function of the particular realization x.[71] The *mutual information* $I(X;Y)$ can then be defined as the mean value of this expression,

$$I(X;Y) \equiv \mathrm{E}[D_{KL}(p(\cdot|x)||p(\cdot))], \tag{H.76}$$

where the expectation is now over possible realization of x, and the mutual information is also necessarily non-negative.[72]

This definition of the mutual information has the attractive feature of being independent of the coordinates used to parameterize the realizations of the variable Y. Suppose that we write $y = \phi(z)$, where $\phi(\cdot)$ is an invertible smooth coordinate transformation between two Euclidean spaces of the same dimension. Then corresponding to the conditional density $p(y|x)$ for any x, there will be a corresponding density function $\tilde{p}(z|x)$ for the random variable Z (which is just the variable Y described using the alternative coordinate system), such that $\tilde{p}(z|x) = p(\phi(z)|x) \cdot D\phi(z)$ for each z, where $D\phi(z)$ is the Jacobian matrix of the coordinate transformation, evaluated at z. It follows that for any z in the support of Z and any x in the support of X,

$$\frac{p(\phi(z)|x)}{p(\phi(z))} = \frac{\tilde{p}(z|x)}{\tilde{p}(z)},$$

[71] The value of this quantity is necessarily non-negative because of Jensen's inequality, owing to the concavity of the logarithm.

[72] Note that this definition — rather than the one often given in terms of the average reduction in the entropy of Y from observing X — has the advantage of remaining well-defined even when the random variable Y has a continuous distribution. See Cover and Thomas (2006) for further discussion.

so that

$$D_{KL}(p(\cdot|x)||p(\cdot)) \;=\; D_{KL}(\tilde{p}(\cdot|x)||\tilde{p}(\cdot))$$

for all x. We thus find that the mutual information $I(X;\, Y)$ will be the same as $I(X;\, Z)$: it is unaffected by a change in the coordinates used to parameterize Y.[73]

We can similarly define the mutual information in a case in which the support of Y is not the entire Euclidean space, because of the existence of redundant coordinates in the parameterization of realizations y. Suppose that all vectors y in the support of Y are of the form $y = \phi(z)$, where $\phi(\cdot)$ is a smooth embedding of some lower-dimensional Euclidean space (the support of Z) into a higher-dimensional Euclidean space. Then the information about the possible realizations of y contained in a realization of x is given by the information that x contains about the possible realizations of z. If the joint distribution of X and Z is such that we can define conditional density functions $\tilde{p}(z|x)$, with $\tilde{p}(z|x) > 0$ for all z and x, and a marginal density function $\tilde{p}(z) > 0$ for all z, then we can define the mutual information between X and Z using (H.76) as above. Since mutual information should be independent of the coordinates used to parameterize the variables, we can use the value of $I(X;\, Z)$ as our definition of $I(X;\, Y)$ in this case as well (even though expression (H.75) is not defined in this case).

In the case of interest in this paper, X and Y are variables with a joint distribution that is multivariate Gaussian. Let us consider first the generic case in which the conditional variance-covariance matrix $\mathrm{var}[Y|x]$ is of full rank. (Note that this matrix will be independent of the realization of x, and so can be written $\mathrm{var}[Y|X]$, to emphasize that only the parameters of the joint distribution matter.) In this case $\mathrm{var}[Y]$ is of full rank as well, and

[73] It is equally unaffected by a change in the coordinates used to parameterize X, though we need not show this here.

for any x and y, the ratio of the density functions satisfies

$$\begin{aligned}\log \frac{p(y|x)}{p(y)} &= -\frac{1}{2}\log\frac{\det(\mathrm{var}[Y|x])}{\det(\mathrm{var}[Y])} \\ &\quad -\frac{1}{2}(y-\mathrm{E}[y|x])'\mathrm{var}[Y|x]^{-1}(y-\mathrm{E}[y|x]) + \frac{1}{2}(y-\mathrm{E}[y])'\mathrm{var}[Y]^{-1}(y-\mathrm{E}[y]).\end{aligned}$$

Hence for any x, we have

$$D_{KL}(x) = -\frac{1}{2}\log\frac{\det(\mathrm{var}[Y|x])}{\det(\mathrm{var}[Y])},$$

and since this will be independent of the realization of x, we similarly will have

$$I(X;\,Y) = -\frac{1}{2}\log\frac{\det(\mathrm{var}[Y|X])}{\det(\mathrm{var}[Y])}. \tag{H.77}$$

One case in which $\mathrm{var}[Y|x]$ will not be of full rank is if $y = Uz$ for some matrix U, where z is a random vector of lower dimension than that of y. (In this case, the rank of $\mathrm{var}[Y|x]$ cannot be greater than the rank of $\mathrm{var}[Z|x]$, which is at most the dimension of z.) Let us suppose that the rank of U is equal to the dimension of z, so that any vector $y = Uz$ is associated with exactly one vector z. In such a case we can, as discussed above, define the mutual information between X and Y to equal the mutual information between X and Z. If $\mathrm{var}[Z|x]$ is of full rank, then we can use the calculations of the previous paragraph to show that

$$I(X;\,Y) = I(X;\,Z) = -\frac{1}{2}\log\frac{\det(\mathrm{var}[Z|X])}{\det(\mathrm{var}[Z])}. \tag{H.78}$$

We turn now to the calculation of the mutual information between the reduced cognitive state $\bar{s}_t$ and the memory state $\bar{m}_{t+1}$, in the case of a law of motion of the form (H.66) for the memory state. We first consider the case in which X_t is of full rank (which, as noted in the text, will be true except when the memory state m_t is completely uninformative). If

$\bar{\Lambda}_t$ and $I - \bar{\Lambda}_t$ are also both matrices of full rank, then

$$\mathrm{var}[\bar{m}_{t+1} \,|\bar{s}_t] \;=\; \Sigma_{\bar{\omega},t+1} \;=\; (I - \bar{\Lambda}_t) X_t \bar{\Lambda}_t'$$

will be of full rank, and

$$\mathrm{var}[\bar{m}_{t+1}] \;=\; \bar{\Lambda}_t X_t \bar{\Lambda}_t' \,+\, \Sigma_{\bar{\omega},t+1} \;=\; X_t \bar{\Lambda}_t'$$

will be of full rank as well. We can then apply (H.77) to obtain

$$I_t \;=\; -\frac{1}{2} \log \frac{\det[(I - \bar{\Lambda}_t) X_t \bar{\Lambda}_t']}{\det[X_t \bar{\Lambda}_t']} \;=\; -\frac{1}{2} \log \det(I - \bar{\Lambda}_t), \tag{H.79}$$

in conformity with equation (3.30) in the text.

In the case that X_t is of full rank, but $\bar{\Lambda}_t$ is varied so that one of its eigenvalues approaches 1 (meaning that $I - \bar{\Lambda}_t$ approaches a singular matrix, while the determinant of $\bar{\Lambda}_t$ remains bounded away from zero), the value of I_t implied by (H.79) grows without bound. It thus makes sense to assign a value of $+\infty$ to the mutual information in the case that $\bar{\Lambda}_t$ is of full rank but $I - \bar{\Lambda}_t$ is not. Note that in this case there is a linear combination of the elements of $\bar{s}_t$ that is revealed with perfect precision by the memory state (since $\Sigma_{\bar{\omega},t+1}$ will be singular), while this linear combination is a continuous random variable with positive variance (since X_t is of full rank). This is not consistent with any finite value for the mutual information (and so cannot represent a feasible memory structure).

Suppose instead that while X_t is of full rank, $\bar{\Lambda}_t$ is only of rank one. In this case, we have shown above that $\bar{\Lambda}_t$ must be of the form (H.71), as a consequence of which $\Sigma_{\bar{\omega},t+1}$ must be given by (H.72). In this case, the memory state can be represented in the form $\bar{m}_{t+1} = X_t v_t \cdot \tilde{m}_{t+1}$, where $\tilde{m}_{t+1}$ is a scalar random variable with law of motion (H.73). This implies that $\mathrm{var}[\tilde{m}_{t+1} \,|s_t] \;=\; \mathrm{var}[\tilde{\omega}_{t+1}] \;=\; \lambda_t(1 - \lambda_t)$, while $\mathrm{var}[\tilde{m}_{t+1}] \;=\; \lambda_t$. In the case

that $0 < \lambda_t < 1$, we can then apply (H.78) to show that

$$I_t \;=\; -\frac{1}{2}\log\frac{\lambda_t(1-\lambda_t)}{\lambda_t} \;=\; -\frac{1}{2}\log(1-\lambda_t), \tag{H.80}$$

Since in this case, $\det(I - \hat{\Lambda}_t) = \det(I - \lambda_t v_t v_t') = 1 - \lambda_t$, result (H.80) is again just what (H.79) would imply, so that (H.79) continues to be correct even though $\bar{\Lambda}_t$ is singular.

If we consider a sequence of matrices of this kind in which λ_t approaches 1, the mutual information (H.80) grows without bound. Thus we can assign the value $+\infty$ to I_t in the case that $\bar{\Lambda}_t$ is a matrix of rank one with $\lambda_t = 1$. Indeed, in this case, the memory state reveals with perfect precision the value of $v_t'\bar{s}_t$, a continuous random variable with positive variance (under the assumption that X_t is of full rank); but this is not possible in the case of any finite bound on mutual information. Hence (H.79) can be applied to this case as well.

Suppose instead that X_t is of full rank, but $\bar{\Lambda}_t = 0$. In this case, the distribution of $\bar{m}_{t+1}$ is independent of the value of s_{t+1}, and the mutual information between these two variables must be zero. This is also what (H.79) would imply, so that (H.79) is correct in this case as well.

Finally, consider the case in which $X_t = X_0$, the only possible case in which X_t is not of full rank. In this case, we have defined $\mathcal{L}(X_0)$ to consist only of matrices of the form (H.71), with the vector v_t given by (3.38). If $\lambda_t = 0$, then the entire matrix $\bar{\Lambda}_t = 0$, and the argument in the previous paragraph again applies. Suppose instead that $\lambda_t > 0$. Just as in the discussion above of the case of a singular transition matrix, the memory state can be represented by a scalar state variable $\tilde{m}_{t+1}$ with law of motion (H.73), and we can apply (H.78) to show that I_t will be given by (H.80). Again this is just what (H.79) would imply, so that (H.79) also yields the correct conclusion when X_t is a singular matrix.

Thus in all cases, (H.79) applies, and the value of I_t depends only on the choice of the transition matrix $\bar{\Lambda}_t$. It follows that for any sequence of transition matrices $\{\bar{\Lambda}_t\} \in \mathcal{L}^{seq}$,

there will be uniquely defined sequences $\{MSE_t, I_t\}$, allowing the objective (3.18) to be evaluated.

8.5 Recursive Determination of the Optimal Memory Structure

We have shown in the text how the optimal memory structure can be characterized if we can find the value function $V(\hat{\sigma}_t^2)$ that satisfies the Bellman equation

$$V(\hat{\sigma}_t^2) \;=\; \min_{\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))} [\alpha \hat{\sigma}_t^2 \;+\; c(I(\bar{\Lambda}_t)) \;+\; \beta V(f(\hat{\sigma}_t^2, \lambda_t, v_t))]. \tag{H.81}$$

Here we establish some properties of the solution to the optimization problem on the right-hand side of (H.81) for an arbitrary function $V \in \mathcal{F}.$, which we can then be used to establish properties of the value function $V(\hat{\sigma}_t^2)$ that solves this equation, and properties of the optimal memory structure.

Monotonicity of the value function

We first show that, for any function V that may be assumed in the problem on the right-hand side of (H.81), the minimum achievable value of the right-hand side is a monotonically increasing function of $\hat{\sigma}_t^2$. This in turn implies that the value function (which must satisfy (H.81)) must be a monotonically increasing function of its argument.

Fix any value function V to be used in the problem on the right-hand side of (H.81), and consider any two possible degrees of uncertainty $\hat{\sigma}_a^2, \hat{\sigma}_b^2$, satisfying

$$0 \;\leq\; \hat{\sigma}_a^2 \;<\; \hat{\sigma}_b^2 \;\leq\; \sigma_0^2. \tag{H.82}$$

Let $\bar{\Lambda}_t \;=\; \bar{\Lambda}_b$ be some element of $\mathcal{L}(X(\hat{\sigma}_b^2))$, and thus a feasible memory structure when $\hat{\sigma}_t^2 = \hat{\sigma}_b^2$, and let us further suppose that $I(\bar{\Lambda}_b) < \infty$, as must be true of an optimal memory structure. We wish to show that we can choose a transition matrix $\bar{\Lambda}_a \in \mathcal{L}(X(\hat{\sigma}_a^2))$ such

that

$$f(\hat{\sigma}_a^2, \bar{\Lambda}_a) \;=\; f(\hat{\sigma}_b^2, \bar{\Lambda}_b), \tag{H.83}$$

and in addition

$$I(\bar{\Lambda}_a) \;\leq\; I(\bar{\Lambda}_b). \tag{H.84}$$

That is, in the case of the smaller degree of uncertainty $\hat{\sigma}_a^2$ in the cognitive state in period t, it is possible to choose a memory structure that implies exactly the same degree of uncertainty in period $t+1$, and hence the same value for $V(\hat{\sigma}_{t+1}^2)$, at no greater an information cost, and thus it is possible to achieve a strictly lower value for the right-hand side of (H.81).

If we can show this for an arbitrary transition matrix $\bar{\Lambda}_b \in \mathcal{L}(X(\hat{\sigma}_b^2))$, then it is also true when $\bar{\Lambda}_b$ is the transition matrix associated with the optimal memory structure (the solution to the problem on the right-hand side of (H.81)) when $\hat{\sigma}_t^2 = \hat{\sigma}_b^2$. This implies that it is possible to achieve a lower value for the right-hand side of (H.81) when $\hat{\sigma}_t^2 = \hat{\sigma}_a^2$ than it is possible to achieve when $\hat{\sigma}_t^2 = \hat{\sigma}_b^2$. Since this must be true for any values of $\hat{\sigma}_a^2, \hat{\sigma}_b^2$ consistent with (H.82), the right-hand side of (H.81) defines a monotonically increasing function of $\hat{\sigma}_t^2$.

To show that such a construction is always possible, let us first consider the case in which $\hat{\sigma}_b^2 = \hat{\sigma}_0^2$, so that the memory state m_t is completely uninformative in case b. In this case, the assumption that $\bar{\Lambda}_b \in \mathcal{L}(X(\hat{\sigma}_b^2)) = \mathcal{L}(X_0)$ requires that

$$\bar{\Lambda}_b \;=\; \lambda_b \frac{ww'}{w'w}$$

for some $0 \leq \lambda_b < 1$.[74] In this case, the memory structure for the following period is equivalent to one in which there is a univariate memory state

$$\tilde{m}_b \;=\; \frac{\lambda_b}{(\Omega + \sigma_y^2)^{1/2}} y_t \;+\; \tilde{\omega}_b, \qquad \tilde{\omega}_b \;\sim\; N(0,\, \lambda_b(1-\lambda_b)).$$

[74]The upper bound is required in order to satisfy the assumption that $I(\bar{\Lambda}_b) < \infty$.

The implied uncertainty in the following period (given the memory state, but before y_{t+1} is observed) is then given by

$$\Sigma_{t+1} \;=\; \Sigma_0 \;-\; \lambda_b(\Omega + \sigma_y^2)ww'. \tag{H.85}$$

Now let $\bar{s}_a$ be the reduced cognitive state in period t, in the case of a more informative memory structure that implies the lower degree of uncertainty $\hat{\sigma}_a^2$, and let $X_a \equiv X(\hat{\sigma}_a^2)$ be the variance of this random vector. In this case, we can choose a memory structure for the following period defined by the transition matrix

$$\bar{\Lambda}_a \;=\; \lambda_b X_a \frac{e_2 e_2'}{\Omega + \sigma_y^2}$$

where $e_2 \equiv [0\ 1]'$. This is a matrix of the form (H.71), and hence an element of $\mathcal{L}(X_a)$. Because $\bar{\Lambda}_a$ is singular, the specified memory structure is equivalent to one in which there is a univariate memory state

$$\tilde{m}_a \;=\; \lambda_b \frac{e_2'\bar{s}_a}{(e_2' X_a e_2)^{1/2}} \;+\; \tilde{\omega}_a, \qquad \tilde{\omega}_a \;\sim\; N(0,\, \lambda_b(1-\lambda_b)).$$

But this means that

$$\tilde{m}_a \;=\; \frac{\lambda_b}{(\Omega + \sigma_y^2)^{1/2}} y_t \;+\; \tilde{\omega}_a, \qquad \tilde{\omega}_a \;\sim\; N(0,\, \lambda_b(1-\lambda_b)).$$

Hence the joint distribution of $(\tilde{m}_a,\, x_{t+1})$ is identical to the joint distribution of $(\tilde{m}_b,\, x_{t+1})$, and the implied uncertainty in the following period given this memory structure is again given by (H.85). Hence the value of $\hat{\sigma}_{t+1}^2$ implied by memory structure a is the same as that implied by memory structure b. This establishes condition (H.83). Moreover, for both memory structures we have the same mutual information,

$$I(\bar{\Lambda}_a) \;=\; I(\bar{\Lambda}_b) \;=\; -\frac{1}{2}\log(1-\lambda_b).$$

This establishes condition (H.84). Hence the value of the right-hand side of (H.81) must be lower when $\hat{\sigma}_t^2 = \hat{\sigma}_a^2$.

Let us next consider the less trivial case in which $0 < \hat{\sigma}_b^2 < \hat{\sigma}_0^2$. Let $\bar{s}_b$ be the reduced cognitive state in period t that implies a degree of uncertainty $\hat{\sigma}_b^2$, and let $X_b \equiv X(\hat{\sigma}_b^2)$ be the variance of this random vector. Let the optimal memory structure for the following period (the solution to the problem on the right-hand side of (H.81)) in this case be

$$\bar{m}_b \;=\; \bar{\Lambda}_b \bar{s}_b \,+\, \bar{\omega}_b, \tag{H.86}$$

where

$$\bar{\Lambda}_b \,\in\, \mathcal{L}(X_b), \qquad \bar{\omega}_b \,\sim\, N(0,\, (I - \bar{\Lambda}_b) X_b \bar{\Lambda}_b').$$

The implied uncertainty in the following period will then be given by

$$\Sigma_{t+1} \;=\; \Sigma_0 \,-\, X_b \bar{\Lambda}_b'. \tag{H.87}$$

Let us consider the memory structure for cognitive state a defined by the transition matrix

$$\bar{\Lambda}_a \;=\; \bar{\Lambda}_b \Gamma \Psi \Gamma^{-1}, \tag{H.88}$$

where Γ is the invertible matrix defined in (H.67), and

$$\Psi \;\equiv\; \begin{bmatrix} \psi & 0 \\ 0 & 1 \end{bmatrix},$$

where $0 < \psi < 1$ is the quantity

$$\psi \;\equiv\; \frac{\hat{\sigma}_0^2 - \hat{\sigma}_b^2}{\hat{\sigma}_0^2 - \hat{\sigma}_a^2}.$$

Note that Ψ is a diagonal matrix, with the property that

$$\Psi \check{X}_a \;=\; \check{X}_a \Psi \;=\; \check{X}_b,$$

using the notation $\check{X}_i \equiv \check{X}(\hat{\sigma}_i^2)$ for $i = a, b$, where $\check{X}(\hat{\sigma}_t^2)$ is the function defined in (H.68). It is first necessary to verify that $\bar{\Lambda}_a \in \mathcal{L}(X_a)$, so that this matrix defines a possible memory structure.

We first show that $\bar{\Lambda}_a X_a = X_a \bar{\Lambda}_a'$. Definition (H.88) implies that

$$\begin{aligned}
\bar{\Lambda}_a X_a &= \bar{\Lambda}_b \Gamma \Psi \Gamma^{-1} X_a \\
&= \bar{\Lambda}_b \Gamma \Psi \check{X}_a \Gamma' \\
&= \bar{\Lambda}_b \Gamma \check{X}_b \Gamma' \\
&= \bar{\Lambda}_b X_b.
\end{aligned}$$

The fact that $\bar{\Lambda}_b \in \mathcal{L}(X_b)$ implies that $\bar{\Lambda}_b X_b$ must be a symmetric matrix; hence $\bar{\Lambda}_a X_a$, which is the same matrix, must also be symmetric. Thus $\bar{\Lambda}_a X_a = X_a \bar{\Lambda}_a'$.

Next, we must also show that $(I - \bar{\Lambda}_a) X_a \bar{\Lambda}_a'$ is a p.s.d. matrix. We first note that $I - \Psi$ is a diagonal matrix with non-negative elements on the diagonal; it follows that $(I - \Psi)\check{X}_b$ is also a diagonal matrix with non-negative elements on the diagonal, and hence p.s.d. From this it follows that

$$\begin{aligned}
\bar{\Lambda}_b \Gamma \cdot (I - \Psi) \check{X}_b \cdot \Gamma' \bar{\Lambda}_b' &= \bar{\Lambda}_b \Gamma (\check{X}_b - \Psi \check{X}_a \Psi) \Gamma' \bar{\Lambda}_b' \\
&= \bar{\Lambda}_b (\Gamma \check{X}_b \Gamma') \bar{\Lambda}_b' \;-\; (\bar{\Lambda}_b \Gamma \Psi \Gamma^{-1})(\Gamma \check{X}_a \Gamma')(\bar{\Lambda}_b \Gamma \Psi \Gamma^{-1})' \\
&= \bar{\Lambda}_b X_b \bar{\Lambda}_b' \;-\; \bar{\Lambda}_a X_a \bar{\Lambda}_a' \\
&= (X_a \bar{\Lambda}_a' \;-\; \bar{\Lambda}_a X_a \bar{\Lambda}_a') \;-\; (X_b \bar{\Lambda}_b' \;-\; \bar{\Lambda}_b X_b \bar{\Lambda}_b') \\
&= (I - \bar{\Lambda}_a) X_a \bar{\Lambda}_a' \;-\; (I - \bar{\Lambda}_b) X_b \bar{\Lambda}_b'
\end{aligned}$$

must be p.s.d. as well. But since the fact that $\bar{\Lambda}_b \in \mathcal{L}(X_b)$ implies that $(I - \bar{\Lambda}_b)X_b\bar{\Lambda}_b'$ must be p.s.d., it follows that $(I - \bar{\Lambda}_a)X_a\bar{\Lambda}_a'$ can be expressed as the sum of two p.s.d. matrices, and so must also be p.s.d. This verifies the second of the conditions required in order to show that $\bar{\Lambda}_a \in \mathcal{L}(X_a)$.

Thus if $\bar{s}_a$ is a reduced cognitive state for period t that implies a degree of uncertainty $\hat{\sigma}_a^2$, a possible memory structure for the following period is

$$\bar{m}_a \;=\; \bar{\Lambda}_a\bar{s}_a \,+\, \bar{\omega}_a, \tag{H.89}$$

where the transition matrix $\bar{\Lambda}_a$ is defined in (H.88), and

$$\bar{\omega}_a \;\sim\; N(0,\, (I - \bar{\Lambda}_a)X_a\bar{\Lambda}_a').$$

The implied uncertainty in the following period will then be given by

$$\Sigma_{t+1} \;=\; \Sigma_0 \,-\, X_a\bar{\Lambda}_a'.$$

This latter matrix is the same as the one in (H.87); it follows that the implied value of $\hat{\sigma}_{t+1}^2$ is also the same as for the memory structure (H.86). Thus we have shown that in the case of the smaller degree of uncertainty $\hat{\sigma}_a^2$, it is possible to choose a memory structure that implies exactly the same degree of uncertainty in period $t+1$ as when the degree of uncertainty in period t is given by the larger quantity $\hat{\sigma}_b^2$.

It remains to be shown that memory structure (H.89) involves no greater information cost than memory structure (H.86). Consider first the case in which the memory state $\bar{m}_b$ is non-degenerate, in the sense that $\mathrm{var}[\bar{m}_b] = X_b\bar{\Lambda}_b'$ is non-singular. It follows that the same must be true of memory state $\bar{m}_a$. Then for either of the two memory structures $i = a, b$

just discussed, (H.79) implies that the mutual information will be given by

$$I_t \;=\; -\frac{1}{2}\log\frac{\det[(I-\bar{\Lambda}_i)X_i\bar{\Lambda}_i']}{\det[X_i\bar{\Lambda}_i']}.$$

We have shown above that the value of the denominator in this expression is the same for $i = a, b$ (and under the assumption that $X_b\bar{\Lambda}_b'$ is non-singular, it must be positive). Hence the relative size of the two mutual informations depends on the relative size of the numerator in the two cases. But we have shown above that $(I-\bar{\Lambda}_a)X_a\bar{\Lambda}_a'$ can be expressed as the sum of $(I-\bar{\Lambda}_b)X_b\bar{\Lambda}_b'$ plus a p.s.d. matrix. Since both of these matrices are also p.s.d., their determinants satisfy

$$\det[(I-\bar{\Lambda}_a)X_a\bar{\Lambda}_a'] \;\geq\; \det[(I-\bar{\Lambda}_b)X_b\bar{\Lambda}_b'] \;>\; 0,$$

where the final inequality is necessary in order for memory structure b to have a finite information cost. It follows that condition (H.84) must hold in this case.

Now suppose instead that $\mathrm{var}[\bar{m}_b]$ is a singular matrix. In the case that the matrix is zero in all elements, $\bar{\Lambda}_b = 0$, and so (H.88) implies that $\bar{\Lambda}_a = 0$ as well. In this case, $\det(I-\bar{\Lambda}_a) = \det(I-\bar{\Lambda}_b) = 1$, so that $I(\bar{\Lambda}_a) = I(\bar{\Lambda}_b) = 0$, and (H.84) is satisfied in this case as well. Thus we need only consider further the case in which $\mathrm{var}[\bar{m}_b]$ is of rank one, which requires that $\bar{\Lambda}_b$ be of rank one as well.

In this case, we can write

$$\bar{\Lambda}_b \;=\; \lambda_b X_b v_b v_b',$$

where $0 < \lambda_b < 1$[75] and v_b is a vector such that $v_b'X_bv_b = 1$. All columns of $\bar{\Lambda}_b$ are multiples of the vector X_bv_b, and as a consequence the unique non-null right eigenvector of $\bar{\Lambda}_b$ is given by X_bv_b, with the associated eigenvalue λ_b. Alternatively, using the orthogonalized

[75] Again, the upper bound is required in order for $I(\bar{\Lambda}_b)$ to be finite.

representation of the cognitive state introduced in section C.4, we can write

$$\Gamma^{-1}\bar{\Lambda}_b\Gamma \;=\; \lambda_b\check{X}_b\check{v}_b\check{v}_b',$$

where we define $\check{v}_b \equiv \Gamma' v_b$, and note that $\check{v}_b'\check{X}_b\check{v}_b = 1$.

Then (H.88) implies that the columns of $\bar{\Lambda}_a$ must also all be multiples of the vector $X_b v_b$. It follows that $\bar{\Lambda}_a$ must also be singular, and that its unique non-null eigenvector must be $X_b v_b$, with an associated eigenvalue

$$\begin{aligned}
\lambda_a &= \lambda_b v_b'\Gamma\Psi\Gamma^{-1}(X_b v_b) \\
&= \lambda_b\check{v}_b'\Psi\check{X}_b\check{v}_b \\
&= \lambda_b(\check{v}_b'\Psi^{1/2})\check{X}_b(\Psi^{1/2}\check{v}_b) \\
&\leq \lambda_b\check{v}_b'\check{X}_b\check{v}_b \;=\; \lambda_b.
\end{aligned}$$

Thus we must have

$$\det(I - \bar{\Lambda}_a) \;=\; (1-\lambda_a) \;\geq\; (1-\lambda_b) \;=\; \det(I-\bar{\Lambda}_b),$$

from which it follows that (H.84) must hold in this case as well.

Thus we have shown that whenever $\hat{\sigma}_a^2, \hat{\sigma}_b^2$ satisfy (H.82), for any memory structure for case b with a finite information cost, it is possible to choose a memory stucture for case a satisfying both (H.83) and (H.84). This means that it must be possible to achieve a lower value for the right-hand side of (H.81) when $\hat{\sigma}_t^2 = \hat{\sigma}_a^2$ than when $\hat{\sigma}_b^2$. This in turn implies that the right-hand side of (H.81) defines a monotonically increasing function of $\hat{\sigma}_t^2$, regardless of the nature of the function $V(\hat{\sigma}_{t+1}^2)$ that is assumed in this optimization problem. Hence the value function $V(\hat{\sigma}_t^2)$ defined by (H.81) must be a monotonically increasing function of its argument.

Optimality of a unidimensional memory state

Here we establish, as stated in the text, that the matrix $\bar{\Lambda}_t$ that solves the problem

$$\min_{\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))} I(\bar{\Lambda}_t) \qquad \text{s.t. } f(\hat{\sigma}_t^2, \bar{\Lambda}_t) \leq \hat{\sigma}_{t+1}^2, \tag{H.90}$$

for given values of $(\hat{\sigma}_t^2, \hat{\sigma}_{t+1}^2)$ is necessarily at most of rank one. As explained in the text, we need only consider the case in which $\hat{\sigma}_t^2 < \hat{\sigma}_0^2$. Given a matrix $\bar{\Lambda}_t$ of rank two that satisfies the constraint in (H.90), we wish to show that we can choose an alternative transition matrix of at most rank one, that also satisfies the constraint, but which achieves a lower value of $I(\bar{\Lambda}_t)$.

We first note that when $\hat{\sigma}_t^2 < \hat{\sigma}_0^2$, $X(\hat{\sigma}_t^2)$ is non-singular. Under the hypothesis that $\bar{\Lambda}_t$ is non-singular, it follows that $X_t \bar{\Lambda}_t'$ is non-singular as well (where we now simply write X_t for $X(\hat{\sigma}_t^2)$), and hence positive definite. Similarly, $\bar{\Lambda}_t X_t \bar{\Lambda}_t'$ must be non-singular and hence positive definite.

Then let the alternative transition matrix be given by

$$\bar{\Lambda}_t^{1D} \;=\; \lambda_t X_t v_t v_t', \tag{H.91}$$

with

$$\lambda_t \;=\; \frac{\delta_{t+1}' \bar{\Lambda}_t X_t \bar{\Lambda}_t' \delta_{t+1}}{\delta_{t+1}' X_t \bar{\Lambda}_t' \delta_{t+1}}, \qquad v_t \;=\; \frac{\bar{\Lambda}_t' \delta_{t+1}}{(\delta_{t+1}' \bar{\Lambda}_t X_t \bar{\Lambda}_t' \delta_{t+1})^{1/2}},$$

where $\delta_{t+1} \equiv e_1 - \gamma_{1,t+1} c$ is the vector introduced in (3.34), and let the matrix $\Sigma_{\bar{\omega},t+1}$ be correspondingly modified in the way specified by (3.27). The fact that $X_t \bar{\Lambda}_t'$ is positive definite implies that the denominator of the expression for λ_t is necessarily positive, so that this quantity is well-defined. Similarly, the fact that $\bar{\Lambda}_t X_t \bar{\Lambda}_t'$ is positive definite implies that the denominator of the expression for v_t is necessarily positive, so that this vector is well-defined as well.

In addition, the fact that (by assumption) $\bar{\Lambda}_t \in \mathcal{L}(X_t)$ implies that $(I - \bar{\Lambda}_t) X_t \bar{\Lambda}_t'$ must be

p.s.d. From this it follows that

$$\delta'_{t+1}(I - \bar{\Lambda}_t)X_t\bar{\Lambda}'_t\delta_{t+1} \;\geq\; 0,$$

and hence that

$$\delta'_{t+1}X_t\bar{\Lambda}'_t\delta_{t+1} \;\geq\; \delta'_{t+1}\bar{\Lambda}_tX_t\bar{\Lambda}'_t\delta_{t+1} \;>\; 0,$$

where the final inequality follows from the fact that $\bar{\Lambda}_tX_t\bar{\Lambda}'_t$ is positive definite. Thus the proposed definition of λ_t satisfies $0 < \lambda_t \leq 1$. One also observes from the definition of v_t that $v'_tX_tv_t = 1$. These conditions suffice to establish that the alternative transition matrix $\bar{\Lambda}_t^{1D}$ is also an element of $\mathcal{L}(X_t)$. That is, it represents a feasible memory structure for period t, given the value of $\hat{\sigma}_t^2$.

This alternative transition matrix corresponds to a memory structure in which $\bar{m}_{t+1} = X_tv_t\tilde{m}_{t+1}$, where $\tilde{m}_{t+1}$ is the unidimensional memory state with law of motion (3.37). From this it follows that

$$\delta'_{t+1}\bar{m}_{t+1} \;=\; \lambda_t\delta'_{t+1}X_tv_tv'_t\bar{s}_t \;+\; \delta'_{t+1}X_tv_t\tilde{\omega}_{t+1}$$

will be a normally distributed random variable, with conditional first and second moments given by

$$\begin{aligned}
\mathrm{E}[\delta'_{t+1}\bar{m}_{t+1}\,|s_t] &= \lambda_t\delta'_{t+1}X_tv_tv'_t\bar{s}_t \\
&= \frac{\delta'_{t+1}\bar{\Lambda}_tX_t\bar{\Lambda}'_t\delta_{t+1}}{\delta'_{t+1}X_t\bar{\Lambda}'_t\delta_{t+1}}\frac{\delta'_{t+1}X_t\bar{\Lambda}'_t\delta_{t+1}\cdot\delta'_{t+1}\bar{\Lambda}_t\bar{s}_t}{\delta'_{t+1}\bar{\Lambda}_tX_t\bar{\Lambda}'_t\delta_{t+1}} \\
&= \delta'_{t+1}\bar{\Lambda}_t\bar{s}_t
\end{aligned}$$

and

$$\begin{aligned}
\text{var}[\delta'_{t+1}\bar{m}_{t+1}\,|s_t] &= \lambda_t(1-\lambda_t)(\delta'_{t+1}X_t v_t)^2 \\
&= (1-\lambda_t)\frac{\delta'_{t+1}\bar{\Lambda}_t X_t \bar{\Lambda}'_t \delta_{t+1}}{\delta'_{t+1}X_t\bar{\Lambda}'_t\delta_{t+1}}\frac{(\delta'_{t+1}X_t\bar{\Lambda}'_t\delta_{t+1})^2}{\delta'_{t+1}\bar{\Lambda}_t X_t\bar{\Lambda}'_t\delta_{t+1}} \\
&= (1-\lambda_t)\delta'_{t+1}X_t\bar{\Lambda}'_t\delta_{t+1} \\
&= \delta'_{t+1}X_t\bar{\Lambda}'_t\delta_{t+1} \;-\; \delta'_{t+1}\bar{\Lambda}_t X_t\bar{\Lambda}'_t\delta_{t+1} \\
&= \delta'_{t+1}[(I-\bar{\Lambda}_t)X_t\bar{\Lambda}'_t]\delta_{t+1} \\
&= \delta'_{t+1}\Sigma_{\bar{\omega}_{t+1}}\delta_{t+1}.
\end{aligned}$$

These are the same conditional mean and variance as in the case of the memory structure specified by the transition matrix $\bar{\Lambda}_t$. Since the optimal estimate $\hat{\mu}_{t+1}$ depends on m_{t+1} only through the value of $\delta'_{t+1}\bar{m}_{t+1}$ (from equation (3.34)), it follows that the conditional distribution $\hat{\mu}_{t+1}|s_t, y_{t+1}$ will be the same under the alternative memory structure. This in turn implies that the variance of $\hat{\mu}_{t+1}$ will be the same, and hence that

$$\hat{\sigma}^2_{t+1} \;=\; \Omega \;-\; \text{var}[\hat{\mu}_{t+1}]$$

will be the same. Thus $\bar{\Lambda}^{1D}_t$ also satisfies the constraint in (H.90).

Next we show that $I(\bar{\Lambda}^{1D}_t)$ must be lower than $I(\bar{\Lambda}_t)$. Let u'_1 and u'_2 be the two left eigenvectors of $\bar{\Lambda}_t$, with associated eigenvalues μ_1 and μ_2 respectively, and let the eigenvectors be normalized so that $u'_i X_t u_i = 1$ for $i = 1, 2$. The corresponding right eigenvectors must then be $X_t u_1$ and $X_t u_2$ respectively. Thus we have

$$\bar{\Lambda}_t X_t u_i \;=\; \mu_i X_t u_i, \qquad u'_i\bar{\Lambda}_t \;=\; \mu_i u'_i,$$

for $i = 1, 2$, and

$$u'_1 X_t u_1 \;=\; u'_2 X_t u_2 \;=\; 1, \qquad u'_1 X_t u_2 \;=\; 0.$$

The vector δ'_{t+1} introduced in (3.34) can be written as a linear combination of the two left eigenvectors,

$$\delta'_{t+1} \;=\; \alpha_1 u'_1 \;+\; \alpha_2 u'_2,$$

for some coefficients α_1, α_2. This implies that

$$\delta'_{t+1} X_t \bar{\Lambda}'_t \delta_{t+1} \;=\; \alpha_1^2 \mu_1 \;+\; \alpha_2^2 \mu_2,$$

$$\delta'_{t+1} \bar{\Lambda}_t X_t \bar{\Lambda}'_t \delta_{t+1} \;=\; \alpha_1^2 \mu_1^2 \;+\; \alpha_2^2 \mu_2^2,$$

and hence that

$$\lambda_t \;=\; \frac{\alpha_1^2 \mu_1}{\alpha_1^2 \mu_1 + \alpha_2^2 \mu_2} \mu_1 \;+\; \frac{\alpha_2^2 \mu_2}{\alpha_1^2 \mu_1 + \alpha_2^2 \mu_2} \mu_2.$$

Thus we see that λ_t must be a convex combination of μ_1 and μ_2.

The fact that $\bar{\Lambda}_t \in \mathcal{L}(X_t)$ requires that both eigenvalues satisfy $0 \leq \mu_i \leq 1$, and the assumption that $\bar{\Lambda}_t$ is non-singular further requires that $\mu_i > 0$ for both. Thus we must have

$$1 - \mu_i \;> (1 - \mu_1)(1 - \mu_2)$$

for both $i = 1, 2$. Since λ_t is a convex combination of μ_1 and μ_2, it follows that

$$1 - \lambda_t \;>\; (1 - \mu_1)(1 - \mu_2).$$

Thus

$$\det(I - \bar{\Lambda}_t^{1D}) \;=\; 1 - \lambda_t \;>\; (1 - \mu_1)(1 - \mu_2) \;=\; \det(I - \bar{\Lambda}_t).$$

Results (H.79) and (H.80) then imply that $I(\bar{\Lambda}_t^{1D}) < I(\bar{\Lambda}_t)$.

Thus $\bar{\Lambda}_t$ cannot be the solution to the optimization problem (H.90). Since this argument can be made in the case of any matrix $\bar{\Lambda}_t \in \mathcal{L}(X_t)$ that is of full rank, we conclude that the optimal transition matrix can be at most of rank one.

The optimal univariate memory state

We turn now to the question of which linear combination of the elements of the reduced cognitive state constitutes the single variable for which it is optimal to retain a noisy record in memory — that is, we wish to characterize the optimal weight vector v_t in (H.73). Here we take as given the value of λ_t (or equivalently, the mutual information between the period t cognitive state and the memory carried into period $t+1$), and solve for the optimal choice of v_t for any given value of λ_t. With this in hand, it will then be possible to characterize an optimal memory structure in terms of the single parameter λ_t.

Given the value of $\hat{\sigma}_t^2$ and the matrix $X_t \equiv \text{var}[\bar{s}_t]$, and taking as given the value of λ_t, we wish to choose v_t so as to minimize $\hat{\sigma}_{t+1}^2$. Note that

$$\hat{\sigma}_{t+1}^2 \;=\; \min_{\xi,\gamma_1} \text{var}[\mu - \xi\tilde{m}_{t+1} - \gamma_1 y_{t+1}].$$

Hence we can write our problem as the choice of ξ, γ_1, and the vector v_t so as to minimize

$$\begin{aligned} f(\hat{\sigma}_t^2, \lambda_t, v_t;\, \xi, \gamma_1) \;&\equiv\; \text{var}[\mu \,-\, \xi(\lambda_t v_t' \bar{s}_t + \tilde{\omega}_{t+1}) \,-\, \gamma_1 y_{t+1}] \\ &=\; \text{var}[\mu \,-\, \xi\lambda_t v_t' \bar{s}_t \,-\, \gamma_1 y_{t+1}] \,+\, \xi^2 \lambda_t (1-\lambda_t), \end{aligned}$$

subject to the constraint that $v_t' X_t v_t = 1$. Note that the solution to this problem will simultaneously determine the optimal choice of v_t (and hence the optimal memory structure, given a choice of λ_t) and the coefficients of the optimal estimate

$$\hat{\mu}_{t+1} \;=\; \xi\tilde{m}_{t+1} \,+\, \gamma_1 y_{t+1} \tag{H.92}$$

based on that memory structure.

We can alternatively define this problem as the choice of a weighting vector $\psi \equiv \xi\lambda_t v_t$ and a Kalman gain γ_1. The values of these quantities suffice to determine the value of the

objective (if we know the values of $\hat{\sigma}_t^2$ and λ_t), since we can reconstruct ξ and v_t from them:

$$v_t \;=\; \frac{\psi}{(\psi' X_t \psi)^{1/2}}, \qquad \xi \;=\; (\psi' X_t \psi)^{1/2} \lambda_t.$$

Moreover, there is no theoretical restriction on the elements of the vector ψ, since the scale factor ξ can be of arbitrary size in the previous formulation of the optimization problem. Thus we can alternatively state our problem as the choice of a weighting vector ψ and a Kalman gain γ_1 to minimize

$$f(\hat{\sigma}_t^2, \lambda_t;\, \psi, \gamma_1) \;=\; \mathrm{var}[\mu \;-\; \psi' \bar{s}_t \;-\; \gamma_1 y_{t+1}] \;+\; \frac{1-\lambda_t}{\lambda_t} \psi' X_t \psi. \tag{H.93}$$

We can write the first term in this objective as

$$\begin{aligned} \mathrm{var}[\mu \;-\; \psi' \bar{s}_t \;-\; \gamma_1 y_{t+1}] \;&=\; \mathrm{var}[(1-(1-\rho)\gamma_1)(\mu - \hat{\mu}_t) - \gamma_1(y_{t+1} - \mu) + (e_1' - \gamma_1 c')\bar{s}_t - \psi' \bar{s}_t] \\ &=\; (e_1' - \gamma_1 c' - \psi') X_t (e_1 - \gamma_1 c - \psi) \;+\; (1-(1-\rho)\gamma_1)^2 \hat{\sigma}_t^2 \;+\; \gamma_1^2 \sigma_\epsilon^2. \end{aligned}$$

Substituting this into (H.93), we see that the objective is a strictly convex quadratic function of ψ and γ_1, for any values of $\hat{\sigma}_t^2$ and λ_t. It follows that the objective has an interior minimum, given by the unique solution to the first-order conditions.

The FOCs for the minimization of (H.93) are given by the linear equations

$$\psi \;=\; \lambda_t (e_1 - \gamma_1 c), \tag{H.94}$$

$$c' X_t (e_1 - \gamma_1 c - \psi) \;+\; (1-\rho)(1-(1-\rho)\gamma_1)\hat{\sigma}_t^2 \;-\; \gamma_1 \sigma_\epsilon^2 \;=\; 0. \tag{H.95}$$

Equation (H.94) already allows one valuable insight: the optimal weight vector v_t is simply a normalized version of the vector δ_{t+1} defined in (3.34). However, this does not yet tell us how to choose v_t, since the vector δ_{t+1} depends on the Kalman gain $\gamma_{1,t+1}$, which depends on the memory structure chosen in period t.

But together equations (H.94)–(H.95) provide a linear system that can be solved for ψ and γ_1, given the values of $\hat{\sigma}_t^2$ and λ_t. We obtain

$$\gamma_{1,t+1} \;=\; \frac{(1-\lambda_t)\Omega + \lambda_t(1-\rho)\hat{\sigma}_t^2}{(1-\lambda_t)(\Omega+\rho^2\sigma_y^2) + \lambda_t(1-\rho)^2\hat{\sigma}_t^2 + \sigma_\epsilon^2} \tag{H.96}$$

as an explicit solution for the Kalman gain. It is worth noting that this implies that

$$0 \;<\; \gamma_{1,t+1} \;<\; \frac{1}{1-\rho}. \tag{H.97}$$

We can then use this solution to evaluate the elements of the vector δ. We obtain

$$\delta_{1,t+1} \;\equiv\; 1-(1-\rho)\gamma_{1,t+1} \;=\; \frac{(1-\lambda_t)\rho(\Omega+\rho\sigma_y^2) + \sigma_\epsilon^2}{(1-\lambda_t)(\Omega+\rho^2\sigma_y^2) + \lambda_t(1-\rho)^2\hat{\sigma}_t^2 + \sigma_\epsilon^2} \;>\; 0,$$

$$\delta_{2,t+1} \;\equiv\; -\rho\gamma_{1,t+1} \;=\; -\frac{(1-\lambda_t)\rho\Omega + \lambda_t\rho(1-\rho)\hat{\sigma}_t^2}{(1-\lambda_t)(\Omega+\rho^2\sigma_y^2) + \lambda_t(1-\rho)^2\hat{\sigma}_t^2 + \sigma_\epsilon^2} \;\leq\; 0.$$

The weight vector v_t is then just a normalized version of δ_{t+1}.

We note that when $\rho = 0$, the optimal weight vector has $v_2 = 0$; that is, the memory state $\tilde{m}_{t+1}$ is just a noisy record of $\hat{\mu}_t$. (This is intuitive, since when the state is i.i.d., and given the estimate $\hat{\mu}_t$ of the mean, the value of y_t provides no information about anything that needs to be estimated or forecasted in period $t+1$ or later.) Instead when $\rho > 0$, we see that the sign of v_2 is necessarily opposite to the sign of v_1: the optimal memory state averages $\hat{\mu}_t$ and y_t with a negative relative weight on y_t.

Given this solution for γ_1, the implied solution for the vector ψ is given by (H.94). Substituting the solutions for γ_1 and ψ into the quadratic objective, we obtain for the minimum possible value of the objective

$$\hat{\sigma}_{t+1}^2 \;=\; (1-\lambda_t)\delta_{t+1}'\Sigma_0\delta_{t+1} + \lambda_t(\delta_{1,t+1})^2\hat{\sigma}_t^2 \;+\; \gamma_{1,t+1}^2\sigma_\epsilon^2. \tag{H.98}$$

This provides an equation for the evolution of the uncertainty measure $\hat{\sigma}_{t+1}^2$, given a choice

each period of λ_t, and using the formulas above for the values of $\gamma_{1,t+1}$ and δ_{t+1}.

8.6 Numerical Solutions

Here we provide further details of the numerical calculations reported in section 3 of the main text.

Dynamics of uncertainty given the path of $\{\lambda_t\}$

We begin by discussing our approach to numerical solution for the law of motion $\eta_{t+1} = \phi(\eta_t; \lambda_t)$ for the scaled uncertainty measure $\{\eta_t\}$, given a path for the memory-sensitivity coefficient $\{\lambda_t\}$. In terms of this rescaled state variable, the law of motion (H.98) becomes

$$\eta_{t+1} \;=\; (1-\lambda_t)(1-\gamma_{1,t+1})^2 K \;+\; (1-\rho^2\lambda_t)\gamma_{1,t+1}^2 \;+\; \lambda_t(1-(1-\rho)\gamma_{1,t+1})^2\eta_t, \qquad \text{(H.99)}$$

and (H.96) becomes

$$\gamma_{1,t+1} \;=\; \frac{(1-\lambda_t)K + (1-\rho)\lambda_t\eta_t}{(1-\lambda_t)(K+\rho^2) + (1-\rho^2) + (1-\rho)^2\lambda_t\eta_t}. \qquad \text{(H.100)}$$

Substitution of (H.100) for $\gamma_{1,t+1}$ in the right-hand side of (H.99) yields an analytical expression for the function $\phi(\eta_t; \lambda_t)$.

This result suffices to allow us to compute the optimal dynamics of the uncertainty measure $\{\eta_t\}$ in the case that the only limit on the complexity of memory is an upper bound $\lambda_t \leq \bar{\lambda} < 1$ each period. We observe from (H.93) that the objective $f(\hat{\sigma}_t^2, \lambda_t;\, \psi, \gamma_1)$ is minimized, for given values of the other parameters, by making λ_t as large as possible. Hence the same is true for the function $f(\hat{\sigma}_t^2, \lambda_t, v_t)$ obtained by minimizing the objective over possible choices of ξ and γ_1. It follows that it will be optimal to choose $\lambda_t = \bar{\lambda}$ each period in the case of this kind of constraint.

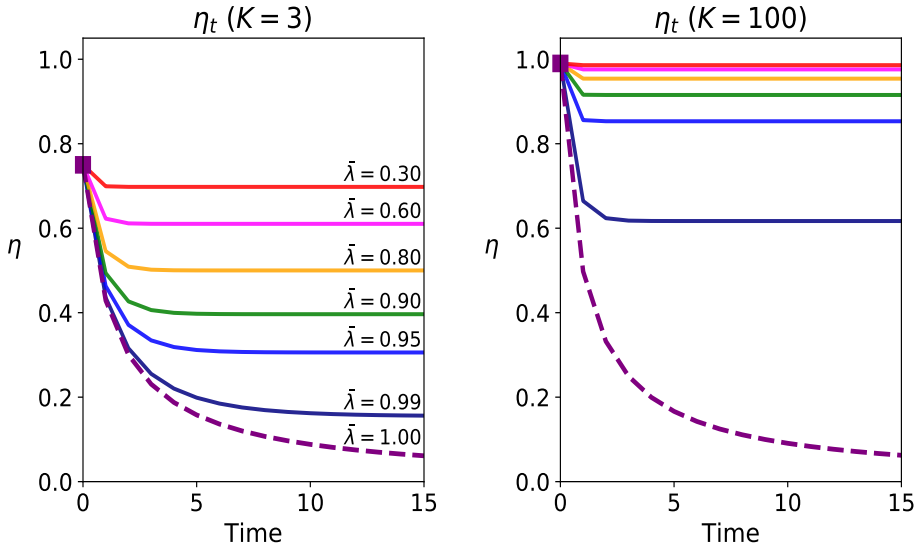


Figure 1.8: The evolution of uncertainty about μ (when $\rho = 0$)

The evolution of scaled uncertainty about μ as the number t of previous (imperfectly remembered) observations grows. Each panel corresponds to a particular value of K (maintaining the assumption that $\rho = 0$, as in Figure 1). Each panel shows the evolution for several different possible values of $\bar{\lambda}$ (color code is the same in both panels).

We thus obtain a nonlinear difference equation

$$\eta_{t+1} \;=\; \phi(\eta_t; \bar{\lambda})$$

for the dynamics of the scaled uncertainty measure. We can iterate this mapping, starting from the initial condition $\eta_0 = K/(K+1)$, to obtain the complete sequence of values $\{\eta_t\}$ for all $t \geq 0$ implied by any given value of $\bar{\lambda}$. This is the method used to compute the dynamic paths shown in Figure 1 in the main text.

Figure 1 shows the dynamics for $\{\eta_t\}$ implied by this solution, for various possible values of $\bar{\lambda}$, in the case that $K = 1$ and $\rho = 0$. Figure 1.8 shows how this graph would be different in the case of two larger values for K (but again assuming $\rho = 0$). A higher value of K (greater prior uncertainty) implies a higher value for the initial value η_0 of our normalized measure of uncertainty (since $\eta_0 = K/(K+1)$). This means that the curves all start higher, the larger the value of K. But the value of K also affects the long-run

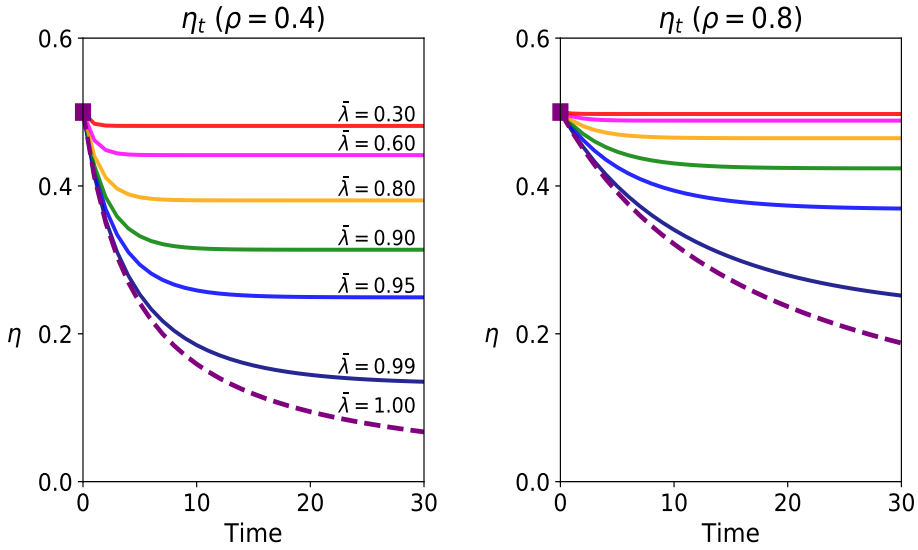


Figure 1.9: The evolution of uncertainty about μ (when $\rho > 0$)

The evolution of scaled uncertainty about μ as the number t of previous (imperfectly remembered) observations grows. Each panel corresponds to a particular value of ρ (maintaining the assumption that $K = 1$, as in Figure 1). Each panel shows the evolution for several different possible values of $\bar{\lambda}$ (color code is the same in both panels).

level of uncertainty η_∞, even though the initial condition becomes irrelevant in the long run. Except when $\bar{\lambda} = 1$ (perfect memory), a higher value of K implies greater long-run uncertainty; and when K is large (as illustrated in the right panel), η_∞ is large (not much below the degree of uncertainty implied by the prior) except in the case of quite high values of $\bar{\lambda}$.

Figure 1.9 similarly shows how Figure 1 would look in the case of two larger values of ρ, but again assuming $K = 1$. We see that for a given degree of prior uncertainty and a given bound on memory precision, the rate at which uncertainty is reduced is slower when the external state is more serially correlated. This is because there are effectively fewer independent observations over a given number of periods when the state is serially correlated. In the case of perfect memory ($\bar{\lambda} = 1$), this affects the speed of learning but not the long-run value $\eta_\infty = 0$ that is eventually reached. Instead, when memory is imperfect, the long-run value η_∞ is also higher when the state is more serially correlated; effectively,

the limited number of recent observations of the state that can be retained in memory reveal less about the value of μ when the state is more serially correlated.

Solving for the value function $\tilde{V}(\eta)$ and policy function $\lambda^*(\eta)$ in the case of a linear information cost

In the case of a linear information cost (or any other cost function with a positive marginal cost of increasing I_t), it is necessary to solve the Bellman equation for the value function $\tilde{V}(\eta)$, in order to determine the optimal dynamics of $\{\lambda_t\}$. Here we explain the methods used to solve this problem in the case of a linear information cost (the results reported in section 3.2).

Once we have solved for the function $\phi(\eta_t; \lambda_t)$, as in the previous subsection, the Bellman equation for the case of a linear information cost can be written

$$\tilde{V}(\eta_t) = \min_{\lambda_t \in [0,1]} \left[\eta_t - \frac{\tilde{\theta}}{2} \log(1 - \lambda_t) + \beta \tilde{V}(\phi(\eta_t; \lambda_t)) \right]. \tag{H.101}$$

We use the value function iteration algorithm to find the value function that is a fixed point of this mapping.

When iterating the mapping to update the value function, we use a grid search method to find the optimal policy function, because the right-hand side of the Bellman equation is in general a non-convex function of the policy variable λ_t (as we illustrate in Figure 1.12 below). We approximate the value function with Chebyshev polynomials. Once the value function has converged, we can use our solution for $\tilde{V}(\eta)$ to solve numerically for the policy function $\lambda^*(\eta)$, the solution to the minimization problem on the right-hand side of (H.101).

This function is graphed for several values of $\tilde{\theta}$ in Figure 1.10, where we maintain the parameter values $K = 1, \rho = 0$ as in Figure 1. When $\tilde{\theta} = 0$ (no cost of memory precision), it is optimal to choose $\lambda_t = 1$ (perfect memory) in all cases. But for any value

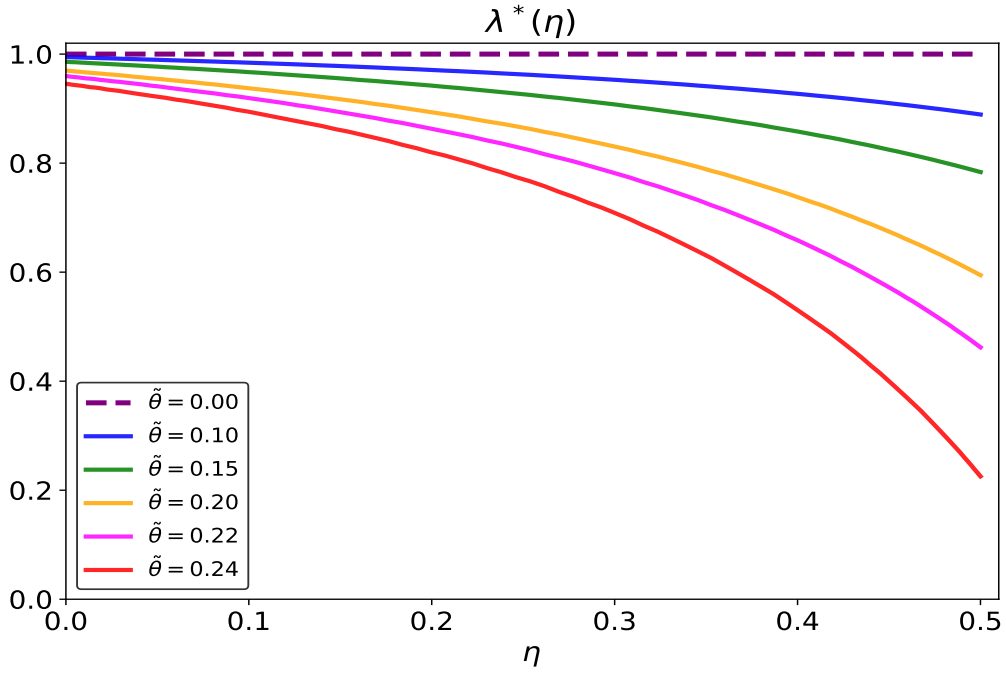


Figure 1.10: The optimal policy function

The optimal policy function $\lambda^*(\eta)$, in the case of progressively larger values for the information cost parameter $\tilde{\theta}$, under the assumption that $K = 1, \rho = 0$.

of η, the optimal $\lambda^*(\eta) < 1$ when $\tilde{\theta} > 0$ (since in this case, perfect memory becomes infinitely costly); furthermore it is lower (memory is more imperfect) the higher is $\tilde{\theta}$. We also see that for any cost parameter $\tilde{\theta} > 0$, the optimal $\lambda^*(\eta)$ is a decreasing function of η. This indicates that the less accurate the information contained in the cognitive state s_t (as indicated by the higher value of η_t), the less information about the cognitive state that it will be optimal to store in memory, when the memory cost can be reduced by storing a less informative record.

The policy function $\lambda_t = \lambda^*(\eta_t)$ together with the law of motion

$$\eta_{t+1} \;=\; \phi(\eta_t; \lambda_t) \tag{H.102}$$

derived in section F.1 can then be solved for the dynamics of the scaled uncertainty $\{\eta_t\}$ for all $t \geq 0$, starting from the initial condition $\eta_0 = K/(K+1)$. The dynamics implied

Figure 4: The dynamics of scaled uncertainty and memory precision

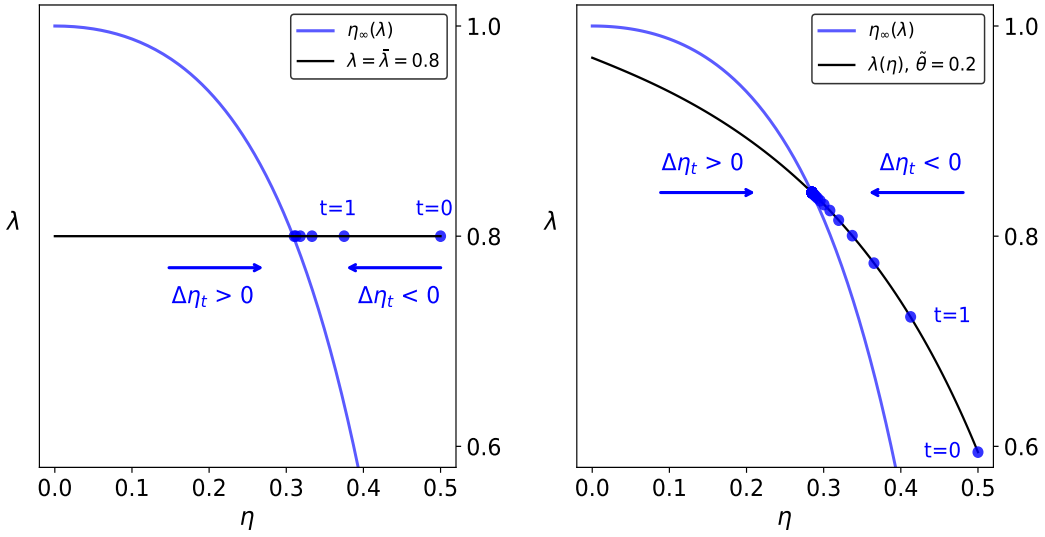


Figure 1.11: The dynamics of scaled uncertainty and memory precision

The dynamics of scaled uncertainty η_t and memory precision λ_t graphed in the phase plane. The left panel gives an alternative graphical presentation of the dynamics plotted in Figure 1 for the case of a fixed upper bound $\bar{\lambda}$ on memory precision. The right panel shows the corresponding dynamics in the case of a linear cost of precision parameterized by $\tilde{\theta}$.

by these equations can be graphed in a phase diagram, as illustrated in Figure 1.11. In the phase diagrams shown in each of the two panels, the value of η_t is indicated on the horizontal axis and the value of λ_t on the vertical axis. Equation (H.102), which holds regardless of the nature of the information cost function and the degree to which the future is discounted, determines a locus $\eta_\infty(\lambda)$, indicating for each value of λ the unique value of η that will be a fixed point of these dynamics if λ_t is held at the value λ. We can further show that whenever $\eta_t < \eta_\infty(\lambda_t)$, the law of motion (H.102) implies that $\eta_{t+1} > \eta_t$, so that uncertainty will increase, while if $\eta_t > \eta_\infty(\lambda_t)$, it implies instead that $\eta_{t+1} < \eta_t$, so that uncertainty will decrease.

The choice of λ_t (and hence the degree to which uncertainty will increase or decrease) is given by the policy function, that depends on the specification of information costs. When there is a fixed upper bound on information (the case discussed in the previous subsection), the policy function is just a horizontal line at the vertical height $\bar{\lambda}$, as shown

in the left panel of the figure.[76] In this case, the values of (η_t, λ_t) in successive periods start at the point $(\eta_0, \bar{\lambda})$, labeled "$t = 0$" in the figure, and then move left along the graph of the policy function (since $\eta_0 > \eta_\infty(\bar{\lambda})$ as shown). They continue to move left along the policy function, with η_t converging asymptotically to $\eta_\infty(\bar{\lambda})$ from above; the stationary long-run values $(\eta_\infty, \lambda_\infty)$ correspond to the point at which the policy function $\lambda = \bar{\lambda}$ intersects the locus of fixed points $\eta_\infty(\lambda)$.

The right-hand panel of the figure shows the corresponding phase-plane dynamics in the less trivial case of a linear cost function for information. In this case, the policy function is instead a downward-sloping curve, as shown in Figure 1.10.[77] Again the values of (η_t, λ_t) in successive periods must always lie on the graph of the policy function; the direction of motion up or down this curve depends on whether the current position lies to the left or right of the locus of fixed points $\eta_\infty(\lambda)$. The initial point (labeled "$t = 0$") is determined as the point on the policy curve with horizontal coordinate given by the initial condition η_0. Since this point lies to the right of the locus of fixed points, the points for successive periods move up and to the left on the policy curve, meaning that λ_t rises as η_t falls.

The scaled uncertainty continues to fall, and the precision of memory continues to rise, until the values (η_t, λ_t) converge to stationary long-run values $(\eta_\infty, \lambda_\infty)$, again corresponding to the point at which the policy function $\lambda^*(\eta)$ intersects the locus of fixed points $\eta_\infty(\lambda)$. Note that convergence is slower in the right panel of the figure than in the left, because in the early periods, when uncertainty is high, a less precise memory is chosen in the linear-cost case, resulting in slower learning from experience.

Different values of $\tilde{\theta}$ correspond to different locations for the policy function $\lambda^*(\eta)$, as shown in Figure 1.10, and hence to different dynamics in the phase plane, converging to

[76] The figure plots the location of this line for the case $\bar{\lambda} = 0.8$. The figure is drawn for parameter values $K = 1, \rho = 0$. Thus the dynamics of uncertainty shown in the figure correspond to the curve labeled $\bar{\lambda} = 0.8$ in Figure 1.

[77] In the figure, the policy function and the implied dynamics are shown for the case in which $\tilde{\theta} = 0.2$, corresponding to one of the intermediate curves shown in Figure 1.10. Again the figure is for the case $K = 1, \rho = 0$, so that the location of the locus of fixed points $\eta_\infty(\lambda)$ and the law of motion (H.102) remain the same as in the left panel.

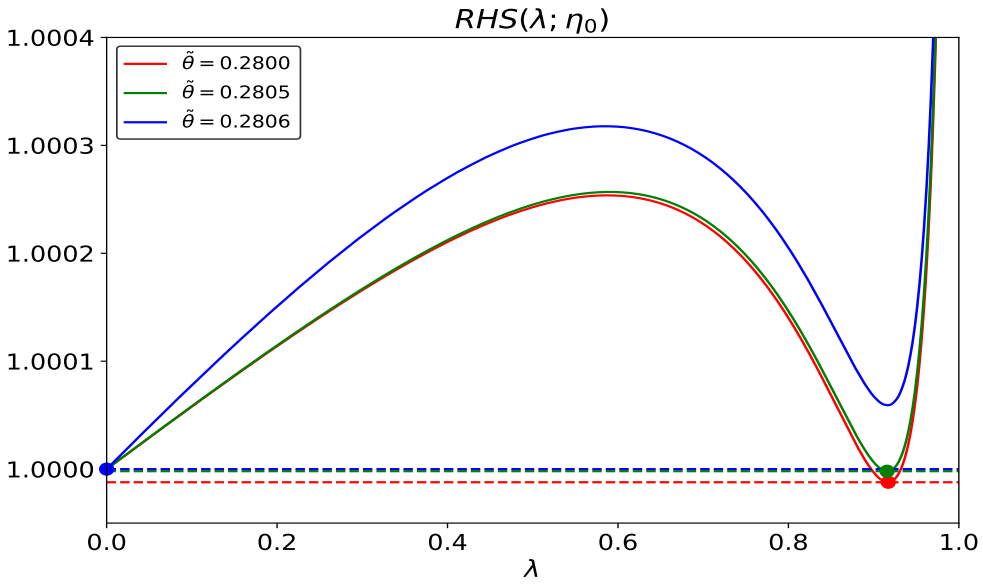


Figure 1.12: The Bellman equation

The objective function $RHS(\lambda_t, \eta_t)$ that is minimized in the Bellman equation, plotted as a function of λ_t for the initial level of uncertainty $\eta_t = \eta_0$. The function is normalized so that the value is 1.0 when $\lambda_t = 0$, and plotted for three nearby values of $\tilde{\theta}$, in the case that $K = 10$. The minimizing value of λ_t jumps discontinuously as $\tilde{\theta}$ passes a value between 0.2800 and 0.2805.

different long-run levels of scaled uncertainty. The dynamics of scaled uncertainty as a function of the number of observations t are shown for progressively larger values of $\tilde{\theta}$ in Figure 3 in the main text, using the same format as in Figure 1.

The possibility of discontinuous solutions

Figure 1.12 illustrates our comment about the possible non-convexity of the optimization problem (H.101). Let $RHS(\lambda_t; \eta_t)$ be the function defined on the right-hand side of (H.101), i.e., the objective of the minimization problem. The figure plots the value of $RHS(\lambda; \eta_0)$, normalized by dividing by the positive constant $RHS(0; \eta_0)$ (so that a value of 1.0 on the vertical axis means that $RHS(\lambda; \eta_0)$ is of exactly the same size as $RHS(0; \eta_0)$). This function is shown for each of three slightly different values of $\tilde{\theta}$, assuming in each case that $K = 10$, as in the right panel of Figure 5 in the text. In the case of each of these curves, a large dot (the same color as the curve) indicates the global minimum of

the function. A horizontal dashed line (also the same color as the corresponding curve) indicates the minimum of $RHS(\lambda;\eta_0)$ — and thus the value of $\tilde{V}(\eta_0)$ — again normalized by dividing by $RHS(\eta_0)$.

The figure shows that for values of $\tilde{\theta}$ in this range, $RHS(\lambda)$ is not a convex function of λ. It is increasing for small enough values of λ, making the choice $\lambda_t = 0$ a local minimum in this case. (This is true for all values of $\tilde{\theta}$ greater than a critical value around 0.15, which explains the existence of the horizontal segment of the connected black curve in the right panel of Figure 5.) However, the function reaches a local maximum, and then decreases for larger values of λ, as the degree to which a larger value of λ_t reduces $\phi(\eta_0;\lambda_t)$ outweighs the increase in the information cost. (A large enough value of K is required for this to occur. A larger value of K increases the sensitivity of the value of $\phi(\eta_0;\lambda)$ to the value of λ; see equation (H.103) below.) For even larger values of λ (values approaching 1), further increases in λ increase the information cost term so sharply that $RHS(\lambda;\eta_0)$ is again decreasing in λ. This means that there is a second local minimum of the objective function, at an interior value of λ. Which of the two local minima represents the global minimum of the function depends on parameter values.

In the case illustrated in the figure, the interior local minimum achieves a lower value of the objective than the choice $\lambda_t = 0$, for all values of $\tilde{\theta}$ less than a critical value that is slightly larger than 0.2805. (As shown in the figure, when $\tilde{\theta}$ = 0.2805, the interior minimum achieves a value of the objective that is quite close to the value $RHS(0;\eta_0)$. However, the value achieved remains slightly smaller: there is a (barely visible) green dashed line, just below the blue dashed line at the normalized value 1.0.) But the normalized value of the objective at the interior minimum increases as $\tilde{\theta}$ is increased, and for a value of $\tilde{\theta}$ only slightly greater than 0.2805, the normalized value becomes greater than 1.0 (which is to say, the interior local minimum is no longer the global minimum of the objective). When this critical value of $\tilde{\theta}$ is passed, the optimal value $\lambda^*(\eta_0)$ jumps discontinuously from the interior local minimum (which is a continuously decreasing function of $\tilde{\theta}$) to the

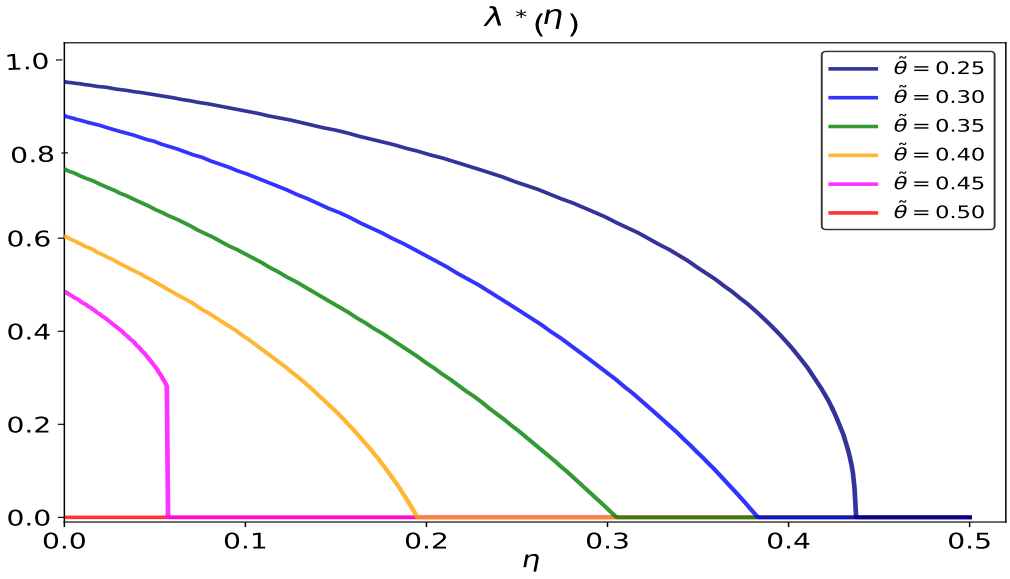


Figure 1.13: The optimal policy function (for a sufficiently large $\tilde{\theta}$)

The optimal policy function $\lambda^*(\eta)$, in the case of progressively larger values for the information cost parameter $\tilde{\theta}$, under the assumption that $K = 1, \rho = 0$. Here we consider values of $\tilde{\theta}$ larger than those shown in Figure 1.10.

value zero. When this happens, the optimal long-run level for the normalized uncertainty measure η_∞ increases discontinuously, from a value on the lower branch of the correspondence shown in the right panel of Figure 5 to the value $\eta_0 = K/K + 1$. For all values of $\tilde{\theta}$ higher than this, it is optimal to choose a completely uninformative memory for all t, so that $\eta_t = \eta_0$ for all t, and hence $\eta_t \to \eta_\infty = \eta_0$.

For larger values of $\tilde{\theta}$ than those considered in Figure 1.10, the optimal policy function $\lambda^*(\eta)$ is equal to zero for all large enough (though still finite) values of η, as illustrated in Figure 1.13. Once $\tilde{\theta}$ is large enough for $\lambda^*(\eta_0)$ to equal zero, the optimal dynamics imply $\eta_t = \eta_0$ for all t, and hence $\eta_\infty = \eta_0 = K/K + 1$, as shown in Figure 5.

The case $\rho = 0$

Additional analytical results are possible in the case that $\rho = 0$ (the external state is an i.i.d. random variable). In this case, the law of motion for the scaled uncertainty measure

(derived in section F.1) simplifies to

$$\eta_{t+1} = 1 - \frac{1}{K + 1 - \lambda_t(K - \eta_t)} \;\equiv\; \phi(\eta_t; \lambda_t). \tag{H.103}$$

In the case of an exogenous upper bound on mutual information, the nonlinear difference equation obtained by setting $\lambda_t = \bar{\lambda}$ in (H.103) is of an especially simple sort. The function on the right-hand side of this equation is a hyperbola, increasing and concave for all $\eta_t > 0$. We easily see that the right-hand side has a positive value when $\eta_t = 0$, and a value less than $K/(K+1)$ when $\eta_t = K/(K+1)$.

Thus for any $0 < \bar{\lambda} < 1$, the function $\phi(\eta_t; \bar{\lambda})$ is an increasing, concave function that is above the diagonal at $\eta_t = 0$ and below the diagonal at $\eta_t = K/(K+1)$. It follows that the function must intersect the diagonal at exactly one point, $\eta_t = \eta_\infty$. We can furthermore give an explicit algebraic solution for this fixed point as the solution to a quadratic equation. Note in particular that it is necessarily strictly positive and strictly less than $K/(K+1)$, and that it is a decreasing function of $\bar{\lambda}$, approaching $K/(K+1)$ as $\bar{\lambda} \to 0$, and approaching 0 as $\bar{\lambda} \to 1$.

On the interval $\eta_\infty < \eta_t \leq K/(K+1)$, the law of motion (H.103) implies that $\eta_\infty < \eta_{t+1} < \eta_t$. Hence when we start from the initial condition $\eta_0 = K/(K+1)$, the implied dynamics must satisfy

$$\eta_0 \;>\; \eta_1 \;>\; \eta_2 \;>\; \eta_3 \;\ldots,$$

a monotonically decreasing sequence. Because the sequence is bounded below by η_∞, it must converge, and it is easily seen that it can only converge to the fixed point η_∞ that we have already calculated. Hence for each possible $\bar{\lambda}$, we obtain a monotonically decreasing, convergent sequence of the kind shown in Figure 1. We can also easily show that the curve must be lower for each value of t, the larger is $\bar{\lambda}$.

We can also obtain additional analytical results in the case of a linear information cost.

The value function satisfies a Bellman equation of the form

$$\tilde{V}(\eta_t) = \min_{\lambda_t} \left[\beta^2 \eta_t - \frac{\tilde{\theta}}{2} \log(1-\lambda) + \beta \tilde{V}\left(\phi(\eta_t; \lambda_t)\right) \right].$$

The first order condition with respect to λ_t is

$$\frac{\tilde{\theta}}{2} \frac{1}{1-\lambda_t} + \beta \tilde{V}'(\eta_{t+1}) \frac{\partial \phi(\eta_t; \lambda_t)}{\partial \lambda_t} = 0. \tag{H.104}$$

And the envelope condition is

$$\tilde{V}'(\eta_t) = \beta^2 + \beta \tilde{V}'(\eta_{t+1}) \frac{\partial \phi(\eta_t; \lambda_t)}{\partial \eta_t}.$$

We can use these two conditions to derive an Euler equation for the dynamics of the scaled uncertainty measure.

Substituting the solution (H.103) for $\phi(\eta_t; \lambda_t)$ and taking the derivative with respect to λ_t, we can rewrite (H.104) as

$$\begin{aligned}
\tilde{V}'(\eta_{t+1}) &= -\frac{\tilde{\theta}}{2\beta} \frac{1}{1-\lambda_t} \left(\frac{\partial \phi(\eta_t; \lambda_t)}{\partial \lambda_t} \right)^{-1} \\
&= -\frac{\tilde{\theta}}{2\beta} \frac{1}{1-\lambda_t} \left(-\frac{(K-\eta_t)}{\left(K+1-\lambda_t(K-\eta_t)\right)^2} \right)^{-1} \\
&= \frac{\tilde{\theta}}{2\beta} \frac{\left(K+1-\lambda_t(K-\eta_t)\right)^2}{(1-\lambda_t)(K-\eta_t)} \\
&= \frac{\tilde{\theta}}{2\beta} \frac{1}{(1-\eta_{t+1})\left(1-(1-\eta_{t+1})(1+\eta_t)\right)},
\end{aligned}$$

where the last equality is derived by again substituting the law of motion (H.103). It follows that if $\eta_t \to \eta_\infty$ in the long run, the stationary solution η_∞ must satisfy

$$\tilde{V}'(\eta_\infty) = \frac{\tilde{\theta}}{2\beta} \frac{1}{(1-\eta_\infty)\eta_\infty^2}. \tag{H.105}$$

Next we rewrite (8.6), again taking the derivative of expression (H.103) for $\tilde{(}\eta_t; \lambda_t)$:

$$\begin{aligned}
\tilde{V}'(\eta_t) &= \beta^2 + \beta\tilde{V}'(\eta_{t+1})\frac{\partial\phi(\eta_t;\lambda_t)}{\partial\eta_t} \\
&= \beta^2 + \beta\tilde{V}'(\eta_{t+1})\frac{\lambda_t}{(K+1-\lambda(K-\eta_t))^2} \\
&= \beta^2 + \beta\tilde{V}'(\eta_{t+1})\frac{\lambda_t}{(1-\eta_{t+1})^{-2}} \\
&= \beta^2 + \beta\tilde{V}'(\eta_{t+1})(1-\eta_{t+1})^2\frac{(K+1)(1-\eta_{t+1})-1}{(K-\eta_t)(1-\eta_{t+1})}.
\end{aligned}$$

It follows that the stationary solution η_∞ must satisfy

$$\tilde{V}'(\eta_\infty) = \beta^2 + \beta\tilde{V}'(\eta_\infty)\frac{(1-\eta_\infty)\left[(K+1)(1-\eta_\infty)-1\right]}{K-\eta_\infty}. \tag{H.106}$$

Moreover, in a stationary solution, the value $\tilde{V}'(\eta_\infty)$ given by (H.105) must also be the value of $\tilde{V}'(\eta_\infty)$ in (H.106). Using (H.105) to substitute for $\tilde{V}'(\eta_\infty)$ in (H.106), we obtain a condition that must be satisfied by η_∞ in any stationary solution with an interior optimum (i.e., a stationary solution in which $0 < \eta_\infty < K/(K+1)$):

$$\tilde{\theta} = 2\beta^3(1-\eta_\infty)\eta_\infty^2\left[1-\beta\frac{(K+1)(1-\eta_\infty)^2-(1-\eta_\infty)}{K-\eta_\infty}\right]^{-1}. \tag{H.107}$$

This is the relationship between $\tilde{\theta}$ and η_∞ that is graphed as a connected black curve in Figure 5. Note that for any value $0 < \eta_\infty < K/(K+1)$, there is a unique $\tilde{\theta} > 0$ consistent with this relationship; but (as shown in the right panel of Figure 5) there may be multiple solutions for η_∞ consistent with a given value of $\tilde{\theta}$.

8.7 Predicted Values for the Quantitative Measures of Forecast Bias

Here we provide further explanation of the numerical results reported in section 4 of the main text.

Long-run stationary fluctuations

From the definition of the univariate memory state $\tilde{m}_{t+1} = \lambda_t v_t' \bar{s}_t + \omega_{t+1}$, we can derive a law of motion for the univariate memory state $\tilde{m}_t$. Using the subscript ∞ for the long-run stationary coefficients, we get

$$\begin{aligned}
\tilde{m}_{t+1} &= \lambda_\infty v_\infty' \bar{s}_t + \tilde{\omega}_{t+1} \\
&= \lambda_\infty v_\infty' \begin{pmatrix} \hat{\mu}_t \\ \\ y_t \end{pmatrix} + \tilde{\omega}_{t+1} \\
&= \lambda_\infty \left[e_1' v_\infty \left\{ (e_1' - \gamma_1 c') m_t + \gamma_1 y_t \right\} + (e_2' v_\infty) y_t \right] + \tilde{\omega}_{t+1} \\
&= \lambda_\infty \left[e_1' v_\infty \left\{ (e_1' - \gamma_1 c') X_\infty v_\infty \tilde{m}_t + \gamma_1 y_t \right\} + (e_2' v_\infty) y_t \right] + \tilde{\omega}_{t+1} \\
&= \rho_m \tilde{m}_t + \rho_{my} y_t + \tilde{\omega}_{t+1}
\end{aligned}$$

where $\rho_m \equiv \lambda_\infty (e_1' v_\infty) \left(e_1' - \gamma_1 c' \right) X_\infty v_\infty$ and $\rho_{my} \equiv \lambda_\infty \left(\gamma_1 + e_2' v_\infty \right)$.

We can evaluate the numerical values of the coefficients defining the long-run dynamics as follows. Equations (H.99)–(H.100) imply that the long-run coefficients $\lambda_\infty, \eta_\infty, \gamma_{1,\infty}$ must satisfy the pair of nonlinear equations

$$\eta_\infty \;=\; \frac{(1-\lambda_\infty)(1-\gamma_{1,\infty})^2 K \;+\; (1-\rho^2\lambda_\infty)\gamma_{1,\infty}^2}{1 \;-\; \lambda_\infty(1-(1-\rho)\gamma_{1,\infty})^2},$$

$$\gamma_{1,\infty} \;=\; \frac{(1-\lambda_\infty)K + (1-\rho)\lambda_\infty\eta_\infty}{(1-\lambda_\infty)(K+\rho^2) + (1-\rho^2) + (1-\rho)^2\lambda_\infty\eta_\infty}.$$

In the case of an exogenous bound on mutual information, we can set $\lambda_\infty = \bar{\lambda}$, in which case these provide two equations to solve for the values of η_∞ and $\gamma_{1,\infty}$. (Note that the relevant solution is the one that satisfies the bounds $0 < \eta_\infty < K/(K+1)$, and that it necessarily also satisfies $0 < \gamma_{1,\infty} < 1/(1-\rho)$.) This allows us to compute the long-run stationary values of the coefficients η and γ_1 plotted for alternative values of $\bar{\lambda}$ in Figure 2.

We have also shown in section E.3 that the optimal weight vector v_t is just a normalized version of the vector $\delta_{t+1} \equiv e_1 - \gamma_{1,t+1}c$. Hence in the long run, this vector must become

$$v_\infty \;=\; \frac{e_1 \;-\; \gamma_{1,\infty}c}{(e_1' - \gamma_{1,\infty}c')X_\infty(e_1 - \gamma_{1,\infty}c)}.$$

In particular, the ratio $v_{2,\infty}/v_{1,\infty}$ (the quantity plotted as "v_∞" in Figure 2) is given by

$$\frac{v_{2,\infty}}{v_{1,\infty}} \;=\; -\frac{\rho\gamma_{1,\infty}}{1 \;-\; (1-\rho)\gamma_{1,\infty}} \;<\; 0.$$

Finally, we observe that the intrinsic persistence coefficient ρ_m defined above must satisfy

$$\begin{aligned}
\rho_m \;&\equiv\; \lambda_\infty v_{1,\infty} \cdot (e_1' - \gamma_{1,\infty}c')X_\infty v_\infty \\
&=\; \lambda_\infty v_{1,\infty} \\
&=\; \lambda_\infty(1 - (1-\rho)\gamma_{1,\infty}).
\end{aligned}$$

This allows us to calculate the other coefficient that is plotted in Figure 2. Note that because the Kalman gain necessarily satisfies the bounds $0 < \gamma_1 < 1/(1-\rho)$, this solution for the intrinsic persistence coefficient implies that

$$0 \;<\; \rho_m \;<\; 1. \tag{H.108}$$

In the long run, we can describe the evolution of the DM's cognitive state using the following system of equations:

$$\begin{aligned}
\tilde{m}_{t+1} &= \rho_m \tilde{m}_t + \rho_{my} y_t + \tilde{\omega}_{t+1} \\
y_{t+1} &= (1-\rho)\mu + \rho y_t + \epsilon_{y,t+1}
\end{aligned}$$

Therefore, we can write it as a VAR(1) system with constant coefficients and Gaussian innovation terms:

$$\begin{pmatrix} \tilde{m}_{t+1} \\ \\ y_{t+1} \end{pmatrix} = \begin{pmatrix} 0 \\ \\ 1-\rho \end{pmatrix} \mu + \begin{pmatrix} \rho_m & \rho_{my} \\ \\ 0 & \rho \end{pmatrix} \begin{pmatrix} \tilde{m}_t \\ \\ y_t \end{pmatrix} + \begin{pmatrix} \tilde{\omega}_{t+1} \\ \\ \epsilon_{y,t+1} \end{pmatrix}$$

Because the two eigenvalues of this vector law of motion are ρ and ρ_m, (H.108) implies that this describes a stationary stochastic process. Hence we can compute stationary long-run values for the second moments of the variables, and use these to define the impulse response functions and predicted regression coefficients reported in the text.

For example, in the case of a fixed per-period bound on mutual information, we can compute the impulse responses for the DM's estimate of μ and her one-quarter-ahead forecast of the external state, as explained in section 3.3. Here we present additional figures, showing what the impulse responses shown in Figure 6 in the text would be like in the case of alternative values of ρ. In Figures 1.14 and 1.15 shown here, each panel corresponds to a different value of ρ, and shows the responses for several different possible values of $\bar{\lambda}$. (As with Figure 6 in the main text, we here assume that $K = 1$.)

Figure 1.14: Impulse responses of the DM's estimate of μ for alternative degrees of persistence ρ of the external state process.

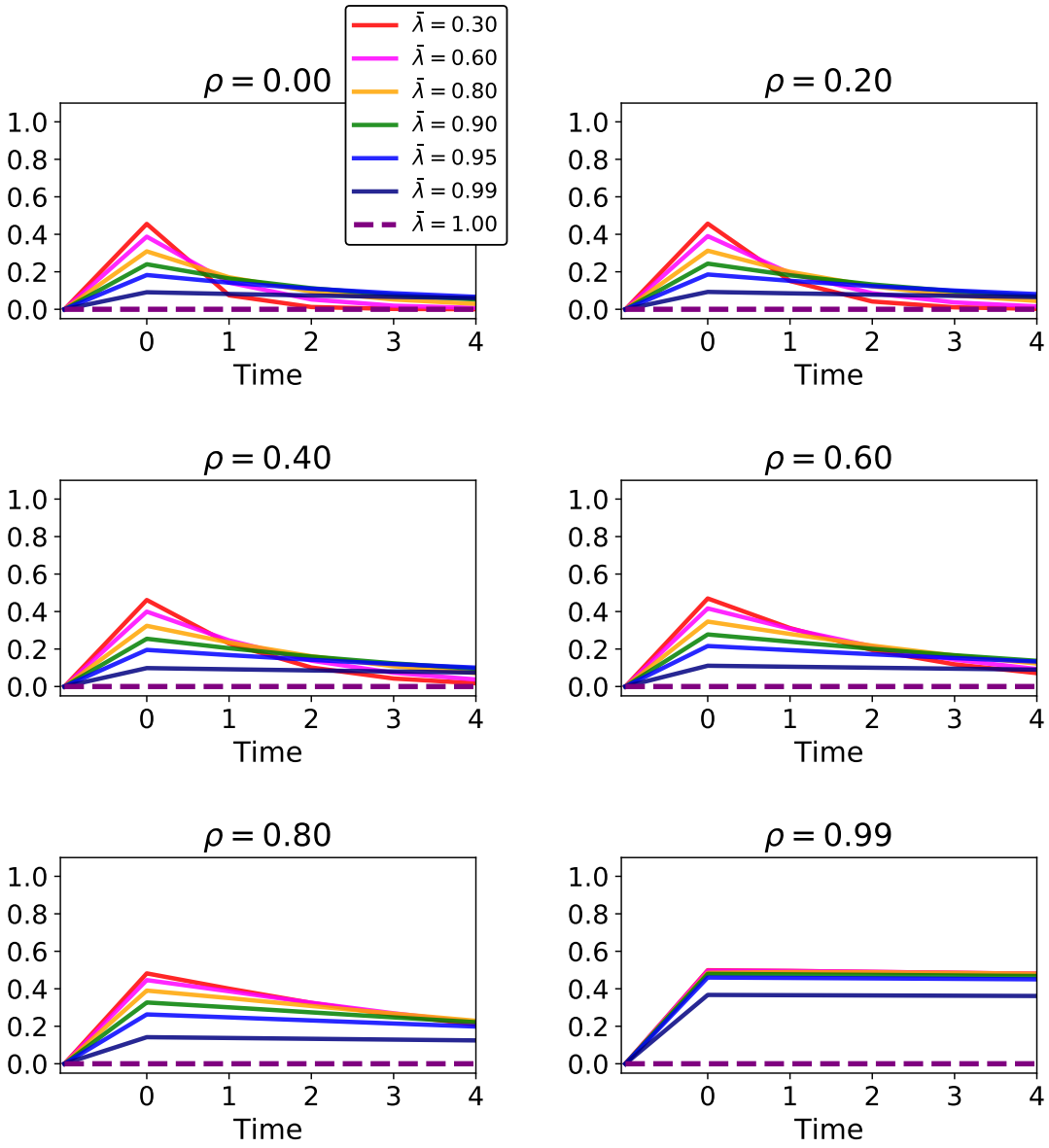
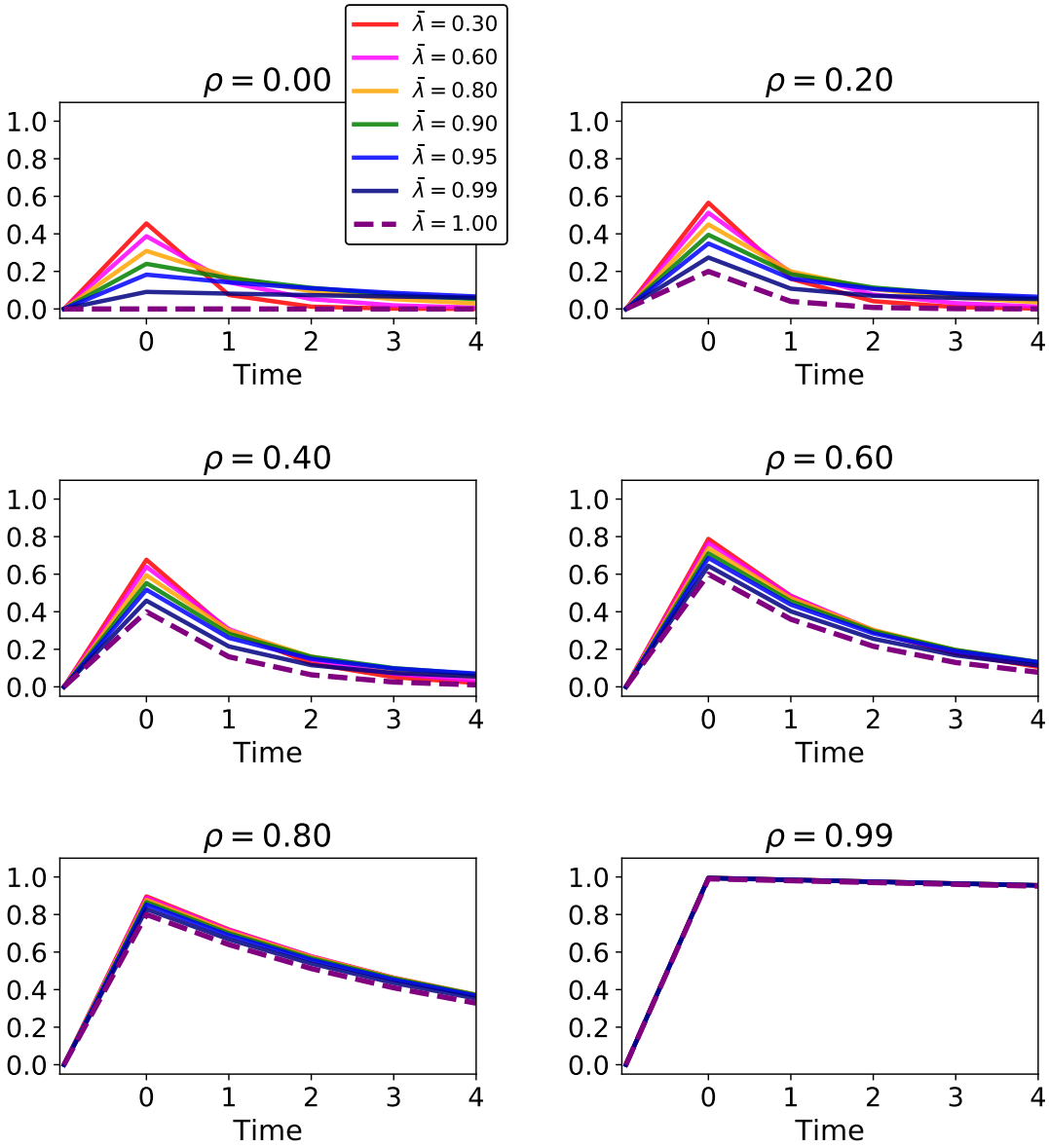


Figure 1.15: Impulse responses of the DM's one-quarter-ahead forecast of the external state for alternative degrees of persistence ρ of the external state process.



Predicted value of the regression coefficient ρ_h^{subj}

Given a long enough series of observations from an environment with a fixed μ, our model yields stationary values for the Kalman gain γ_1 and for the amplitude of fluctuations in the memory state $var[\bar{m}_t]$. We can then compute the values of the following long-run conditional second moments:

$$\begin{aligned} var[\bar{m}_t|\mu] &= var[\bar{m}_t] - cov[\bar{m}_t, \mu] var[\mu]^{-1} cov[\mu, \bar{m}_t] \\ &= var[\bar{m}_t] - cov[\bar{m}_t, x_t] e_1 var[\mu]^{-1} e_1' cov[x_t, \bar{m}_t] \\ &= var[\bar{m}_t] - \frac{1}{var[\mu]} var[\bar{m}_t] e_1 e_1' var[\bar{m}_t] \end{aligned}$$

$$\begin{aligned} cov[\hat{\mu}_t, y_t|\mu] &= cov[(e_1' - \gamma_1 c')\bar{m}_t + \gamma_1 y_t, y_t|\mu] \\ &= (e_1' - \gamma_1 c') cov[\bar{m}_t, y_t|\mu] + \gamma_1 var[y_t|\mu] \\ &= (e_1' - \gamma_1 c') var[\bar{m}_t|\mu] c + \gamma_1 var[y_t|\mu] \end{aligned}$$

$$\begin{aligned} var[\hat{\mu}_t|\mu] &= var[(e_1' - \gamma_1 c')\bar{m}_t + \gamma_1 y_t|\mu] \\ &= (e_1' - \gamma_1 c') var[\bar{m}_t|\mu](e_1 - \gamma_1 c) + \gamma_1^2 var[y_t|\mu] + 2\gamma_1 (e_1' - \gamma_1 c') cov[\bar{m}_t, y_t|\mu] \\ &= (e_1' - \gamma_1 c') var[\bar{m}_t|\mu](e_1 - \gamma_1 c) + \gamma_1^2 var[y_t|\mu] + 2\gamma_1 (e_1' - \gamma_1 c') var[\bar{m}_t|\mu] c \end{aligned}$$

In order to write the dynamics of the model in terms of scale-invariant quantities, we divide each second moment by $var[y_t|\mu] = \sigma_y^2$. Thus we can write

$$\begin{aligned} \frac{var[\bar{m}_t|\mu]}{var[y_t|\mu]} &= \tilde{\Sigma}_{\bar{m}} - \frac{1}{K} \tilde{\Sigma}_{\bar{m}} e_1 e_1' \tilde{\Sigma}_{\bar{m}} \\ \frac{cov[\hat{\mu}_t, y_t|\mu]}{var[y_t|\mu]} &= (e_1' - \gamma_1 c') \frac{var[\bar{m}_t|\mu]}{var[y_t|\mu]} c + \gamma_1 \\ \frac{var[\hat{\mu}_t|\mu]}{var[y_t|\mu]} &= (e_1' - \gamma_1 c') \frac{var[\bar{m}_t|\mu]}{var[y_t|\mu]} (e_1 - \gamma_1 c) + \gamma_1^2 + 2\gamma_1 (e_1' - \gamma_1 c') \frac{var[\bar{m}_t|\mu]}{var[y_t|\mu]} c, \end{aligned}$$

using the notation $\tilde{\Sigma}_{\bar{m}} \equiv \mathrm{var}[\bar{m}_t]/\sigma_y^2$.

We now wish to calculate the predicted asymptotic value of the regression coefficient

$$\rho_h^{subj} \equiv \frac{cov[\hat{y}_{t+h|t}, y_t|\mu]}{var[y_t|\mu]}$$

where $\hat{y}_{t+h|t} \equiv E[y_{t+h}|\bar{m}_t, y_t]$. From

$$\begin{aligned} cov[\hat{y}_{t+h|t}, y_t|\mu] &= cov[(1-\rho^h)\hat{\mu}_t + \rho^h y_t, y_t|\mu] \\ &= (1-\rho^h)cov[\hat{\mu}_t, y_t|\mu] + \rho^h var[y_t|\mu], \end{aligned}$$

where $\hat{\mu}_t \equiv E[\mu|\bar{m}_t, y_t]$, we can then compute

$$\begin{aligned} \rho_h^{subj} &= (1-\rho^h)\frac{cov[\hat{\mu}_t, y_t|\mu]}{var[y_t|\mu]} + \rho^h \\ &= (1-\rho^h)\left[(e_1' - \gamma_1 c')\left(\tilde{\Sigma}_{\bar{m}} - \frac{1}{K}\tilde{\Sigma}_{\bar{m}} e_1 e_1' \tilde{\Sigma}_{\bar{m}}\right) c + \gamma_1\right] + \rho^h. \end{aligned}$$

These are the coefficients whose values are plotted against the value of $\rho_h = \rho^h$ in Figure 7.

Chapter 2: Inflation Surprises and Perception of Inflation Risks

with Miguel Acosta[1]

1 Introduction

Perception of risk is essential in many economic decision-making processes. If households and firms are concerned that extreme events are more likely to arise, they hold off investing in durable goods. In financial markets, market participants assess how much risk an underlying cash flow carries, which is particularly explicit in the pricing of financial contracts, such as variance swaps and inflation swaps.

This paper studies how people perceive risk differently from the optimal Bayesian inference. In particular, we suppose that people are subject to memory friction, and thus risk perception is not accurately based on all available past data. Our model predicts that perceived risk perpetually fluctuates even when there is no change in the true level of risk. Moreover, following large variations of the underlying process, people view extreme realizations are more likely in the future. We validate this model prediction using the inflation forecast data of professional forecasters.

We consider a case where the variance of a random variable is unknown. To clarify the intuition, we consider a random variable following a Gaussian distribution whose mean and variance are constant. We further suppose that decision-maker (DM) knows the mean of the process. Thus, the history of small and large deviations from the mean indicates how dispersed the random variable is. If DM learns about the variance by observing such realized variations, the extent of surprises is informative to gauge the variance, not whether they are old or new.

We propose a learning model where DM's perception is based on noisy memory. We

[1]Division of International Finance, Federal Reserve Board

suppose that perception is not accurately based on past data. Instead, it is based on a somewhat noisy recollection of past data. The essence of our model is to theorize a stock of memory as noisy summary statistics. We first observe that lower dimensional statistics summarize past data informative about the underlying variance. We then suppose that this statistic is only available with noise in the subsequent periods. This formulation nests the perfect memory benchmark, as the extent of the noise is characterized by a single parameter. While we call this model of noisy memory, it does not necessarily mean that DM forgets past data. We formalize the friction in accurately utilizing the past data, which could capture imperfections in information processing other than forgetfulness.

We show that DM overweights recent surprises compared to the perfect memory benchmark. If a statistician were to infer the value of variance from the history of past realizations, then each realized variation carries an equal weight. We show that this is not the case if DM's perception is based on noisy memory. Instead of equal weight, DM puts disproportionately high weight on recent experiences. Furthermore, weights on older data are stochastic and exponentially decaying on average. This means that the sequence of observation matters. Suppose a long period of stable realizations is followed by extreme events. After experiencing this sequence of episodes, a statistician and DM hold different views about the variance. DM views that unlikely events are more likely than what a statistician thinks.

Importantly, such over-extrapolation persists even in the long run. If learning with perfect memory, we expect that DM will eventually learn about the variance in the long run. As learning opportunities accumulate, a new piece of information adds less information and eventually adds close to zero information. Thus, a new experience will stop affecting one's views about the variance. This, again, is not the case if perception is based on noisy memory. We show that DM is perpetually uncertain about the exact value of the variance since knowledge is accumulated with noise. Thus, DM continues to learn even in the long run, which means that DM's risk perception fluctuates and extrapolates from recent ex-

perience. Even in the long run, DM perceives extreme realizations are more likely in the future when the current realization is unusual.

We test the main prediction of the model using the survey data of the US professional forecasters. We find evidence supportive of our theory. When forecasters are surprised by recent inflation data, they become more uncertain about inflation in the future. This positive relationship between the size of inflation surprises and perceived inflation risk holds not only during times of substantial volatility—as in the late 1970s—but also during less-volatile periods. Limitations in the survey's design do not allow us to study the most-recent inflation, but our findings and model suggest that it was likely followed by an increase in the perceived inflation risk.

The proposed model offers a different explanation for fluctuating risk perception compared to a hypothesis that the variance is stochastic. When DM learns about a stochastically-varying variance, the resulting risk perception is similar to ours in that risk perception fluctuates and extrapolates from recent surprises. In this case, older data become obsolete as they are less informative about current variance. Accordingly, fluctuation in risk perception is tightly controlled by the rate at which variance shifts. While this is a plausible explanation, past literature often finds it hard to justify the variation in risk perception with the realistic degree of stochastic movement in the variance process (e.g., see Bakshi and Skoulakis (2010)). In our paper, we provide an alternative theory of fluctuating risk perception even when the volatility of the variance is negligible, in our case, zero.

2 Model

Environment A decision-maker (DM) forecasts a future realization of a variable y_t that follows a normal distribution with mean μ and precision ω:

$$y_t \sim \mathcal{N}\left(\mu,\, \omega^{-1}\right). \tag{B.1}$$

While the DM knows this variable is drawn from a normal distribution, the exact data-generating process—parameterized by μ and ω—is unknown. Our primary interest in this paper is on the DM's perception of the volatility of y_t, which we refer to as perceived volatility. To focus on how beliefs about $\frac{1}{\omega}$ are formed, we start with a more straightforward setup where the DM has a correct understanding of the mean μ. Thus, the DM needs to learn about one parameter: ω.

We assume that the following Gamma distribution describes the DM's prior belief about ω:

$$\omega \sim \Gamma(\alpha,\, \beta)\,.$$

This prior belief ensures that the DM's belief about ω inferred from the history of y_t again follows a Gamma distribution. We define perceived volatility as $E_t\left[\omega^{-1}\right]$, where the expectation is conditional on information available to DM at time t as detailed below.

Available information The DM is Bayesian and thus forms their beliefs about the model's parameters using a combination of their prior information and any new information available to them. The DM learns about the model parameters by observing the realizations of y_t. We make the conventional assumption that at each period, y_t is realized, and the DM observes this realization without error.

Where this paper differs from conventional Bayesian learning is in the use of prior information. This paper studies how beliefs are formed when they are *not* accurately based on all past y_t: The DM's beliefs about the parameters are not described as $\omega|\, y_1, \cdots, y_t$. Instead, past data is less accurately incorporated into the DM's current belief. To clarify this, we describe that a statistician's views are based on "perfect memory," while the DM's views are based on "noisy memory." In the following section, we discuss how noisy memory is modeled.

2.1 Learning with Perfect Memory

In this section, we consider how a statistician's views are formed. This will serve as the full information limit of our model. Suppose the $\mathcal{Y}^t$ denotes the history of y_t, that is, $\mathcal{Y}^t \equiv \{y_1, \cdots, y_t\}$. A statistician accurately infers the value of ω from the history of $\mathcal{Y}^t$.

It is well understood that the conditional distribution $\omega|\,\mathcal{Y}^t$ follows a Gamma distribution and that the distribution evolves in a recursive formula. Here, we summarize this point and introduce a notation. At any time t, we can describe the *prior* belief about ω at time t as

$$\omega|\,\mathcal{Y}^{t-1} \sim \Gamma\left(\alpha + \alpha^*_{t-1},\, \beta + \beta^*_{t-1}\right).$$

Then, an observation of y_t leads DM to have the *posterior* belief about ω as follows.

$$\omega|\,\mathcal{Y}^{t} \sim \Gamma\left(\alpha + \alpha^*_{t} + \frac{1}{2},\, \beta + \beta^*_{t}\right)$$

where α^*_t and β^*_t recursively evolve according to

$$\alpha^*_t = \alpha^*_{t-1} + \frac{1}{2} \tag{B.2a}$$

$$\beta^*_t = \beta^*_{t-1} + \frac{1}{2}\left(y_t - \mu\right)^2. \tag{B.2b}$$

Thus, the distribution $\omega|\,\mathcal{Y}^t$ is completely summarized by the two variables α^*_t and β^*_t. We can see that α^*_t increases incrementally independent of the realization of y_t, while β^*_t reflects how far off the realized y_t is from its expected value μ. Furthermore, the sequence of $\{\alpha^*_\tau, \beta^*_\tau\}^t_{\tau=0}$ is described below.

$$\alpha^*_t = \frac{t}{2} \tag{B.3a}$$

$$\beta^*_t = \frac{1}{2}\sum_{\tau=1}^{t}\left(y_\tau - \mu\right)^2 \tag{B.3b}$$

starting from the initial conditions $\alpha_0 = 0$ and $\beta_0 = 0$. Thus, α_t^* stores the length of the learning experiences, while β_t^* stores the stock of experienced surprises until time t.

The DM's perceived volatility at time t is given by $E[\omega^{-1}|\,\mathcal{Y}^t] = \frac{\alpha+\alpha_t^*}{\beta+\beta^*}$. Thus,

$$E[\omega^{-1}\big|\,\mathcal{Y}^t] = \frac{\beta + \frac{1}{2}\sum_{\tau=1}^{t}(y_\tau - \mu)^2}{\alpha + \frac{t}{2} - 1} \tag{B.4}$$

from equations (B.3). We define the weight on the variation realized in time τ as

$$s_\tau^* = \frac{\frac{1}{2}}{\alpha + \frac{t}{2} - 1}, \quad \tau \leq t. \tag{B.5}$$

As expected, we observe that all previously realized variation receives the same weight. That is, $s_\tau^* = s_t^*$ for all $\tau \leq t$. This means that all experiences have the same influence on DM's current perception of risk, regardless of how recent the experience is.

We can furthermore see that $s_t^* \to 0$ as $t \to \infty$. That is, recent observations will influence DM's perception of ω^{-1} less and less over time. By the law of large numbers, it is also straightforward to see that the estimate (B.4) converges to a true ω^{-1} as learning opportunities accumulate.

2.2 Learning with Noisy Memory

This section considers how the DM's views are formed based on noisy memory. In this case, the DM's beliefs about ω are not accurately based on the history of y_t. Thus, the DM's beliefs less accurately incorporate information from past data than a statistician's beliefs.

We propose a succinct and intuitive theory to introduce memory frictions. Following the work of Azeredo da Silveira et al. (2020) and Sung (2022), our model of memory will posit that the DM's knowledge about the variance of y_t is distilled down to a lower-dimensional set of sufficient statistics. This is motivated by the notion that the DM's beliefs are characterized by parametric distributions (here, the Gamma distribution), then those beliefs are completely summarized by the distributions parameters (in the case of the

Gamma distribution, this means just two variables). Thus, these parameters summarize the knowledge about the history of past data realizations relevant for inferring the level of ω without losing any information.

For example, from the previous section, α_t^* and β_t^* completely describe the posterior distribution. Furthermore, β_t^* captures information from past realizations, as the other parameter α_t^* only depends on the number of observations. Thus, the inference based on β_t^* coincides with the statistician's view: It is sufficient for the DM to base her forecasts on β_t^* to replicate the perfect memory views of a statistician. Lemma 1 summarizes this point.

Lemma 1. *Suppose a posterior belief about ω is given as $\omega|\, \mathcal{Y}^t \sim \Gamma\left(\alpha + \alpha_t^*,\, \beta + \beta_t^*\right)$. Then, it must be that $\omega|\, \beta_t \sim \Gamma\left(\alpha + \alpha_t^*,\, \beta + \beta_t^*\right)$.*

Proof. The lemma follows from $\beta_t^*|\, \omega \sim \Gamma\left(\alpha_t^*,\, \omega\right)$. □

Inspired by the fact that β_t^* fully captures knowledge about ω in the perfect memory case, we suppose that the DM's views are based on noisy memory about this sufficient statistic. Thus, we propose that the DM's knowledge about ω until the beginning of time t is stored in a one-dimensional "memory variable" m_{t+1}, which evolves according to the following definition.

Definition 1. *If the time-t posterior belief about ω is given as $\omega|\, m_t, y_t \sim \Gamma\left(\alpha + \alpha_t,\, \beta + \beta_t\right)$, then memory variable m_{t+1} is defined as*

$$m_{t+1} = \beta_t\, u_{t+1} \tag{B.6}$$

where u_{t+1} is an i.i.d. draw from the Beta distribution $\mathcal{B}\left(\lambda\, \alpha_t,\, (1-\lambda)\, \alpha_t\right)$ for a given scalar $\lambda \in (0, 1)$.

A single parameter λ describes the extent of noisy memory. We can see this from the noise term u_{t+1}. Note that $E[u_{t+1}] = \lambda$ and $V[u_{t+1}] = \frac{\lambda(1-\lambda)}{\alpha+1}$. Thus, the perfect memory case is nested as $\lambda \to 1$ since $m_{t+1} = \beta_t$, in which case the DM's views coincide with a

statistician who accurately uses all available past y_t. On the other hand, the formulation above also incorporates the case in which the memory variable m_{t+1} carries no information about ω. When $\lambda \to 0$, we can see that $m_{t+1} = 0$. Thus, m_{t+1} contains no information about previous realizations that are informative about the variability of y_t. In other words, we have that $\omega | \, m_{t+1} \sim \Gamma(\alpha, \, \beta)$ as m_{t+1} does not improve DM's initial prior about ω.

Given equation (B.6), the DM's beliefs about ω evolve as described in Proposition 1. Importantly, the proposed memory structure allows the prior belief about ω to be the conjugate prior, as was the case in the perfect memory benchmark. Thus, the DM's prior and posterior beliefs about ω are described as Gamma distributions.

Proposition 1. *If memory is formed according to (B.6), then the time-t prior belief can be expressed as*

$$\omega | \, m_t \sim \Gamma(\alpha + \alpha_t^m, \, \beta + \beta_t^m)$$

The posterior belief, after observing y_t, evolves to

$$\omega | \, m_t, y_t \sim \Gamma(\alpha + \alpha_t, \, \beta + \beta_t)$$

where $\alpha_t = \alpha_t^m + \frac{1}{2}$ and $\beta_t = \beta_t^m + \frac{1}{2}(y_t - \mu)^2$. Finally, the beginning-of-period prior at $t+1$ is derived as

$$\omega | \, m_{t+1} \sim \Gamma\left(\alpha + \alpha_{t+1}^m, \, \beta + \beta_{t+1}^m\right)$$

where $\alpha_{t+1}^m = \lambda \, \alpha_t$ and $\beta_{t+1}^m = m_{t+1} = \beta_t \, u_{t+1}$.

Proof. See Appendix 5.1. □

As Proposition 1 suggests, DM's posterior belief about ω follows a Gamma distribution

specified by two variables, α_t and β_t. They evolve recursively as follows.

$$\alpha_t = \lambda\,\alpha_{t-1} + \frac{1}{2} \tag{B.7a}$$

$$\beta_t = \beta_{t-1}\,u_t + \frac{1}{2}\,(y_t - \mu)^2 \tag{B.7b}$$

where u_t is randomly drawn from $\mathcal{B}\,(\lambda\,\alpha_{t-1},\,(1-\lambda)\,\alpha_{t-1})$. Compared to the perfect memory counterpart (B.2), there are important differences in how α_t and β_t evolve. First, while the evolution of α_t from α_{t-1} is still independent of the realization of y_t, the change is somewhat attenuated since $\lambda \in (0,1)$. Second, while the fluctuation in β_t is still tied to the realized variation $(y_t - \mu)^2$, there is a second source of fluctuation: The stochastic term u_t is independent of y_t and multiplies β_{t-1}. Thus, the experience of past surprises summarized in β_{t-1} influences β_t in a stochastic manner.

By recursively iterating (B.7), we can describe α_t and β_t as follows.

$$\alpha_t = \frac{1}{2} + \frac{\lambda}{2} + \cdots + \frac{\lambda^{t-1}}{2} = \frac{1}{2}\,\frac{1-\lambda^t}{1-\lambda} \tag{B.8a}$$

$$\beta_t = \frac{1}{2}\sum_{\tau=1}^{t}\,(y_\tau - \mu)^2\;\bar{u}_{t,\tau} \tag{B.8b}$$

starting from the initial points $\alpha_0 = \beta_0 = 0$. [2] The term $\bar{u}_{t,\tau}$ is defined as $\Pi_{i=\tau+1}^{t} u_i$ for $\tau \leq t-1$ and $\bar{u}_{t,t} = 1$. The equations (B.8) summarize how limited memory affects the DM's learning. Note first that when $\lambda = 1$, α_t and β_t converge to their perfect-memory counterpart α_t^* and β_t^* described in (B.3). However, when memory is limited, α_t does not linearly increase with the length of learning opportunities; α_t converges to a positive scalar as $t \to \infty$. Also, β_t is not simply summing past realized variations. Previous surprises are multiplied by idiosyncratic noise, and the noise is cumulative, as captured by $\bar{u}_{t,\tau}$. Therefore, older surprises have a less precise impact on the DM's estimate of ω.

[2] At $t = 1$, we assume there is no memory to be inherited. The posterior is then $\omega|\,y_1$ which is a Gamma distribution.

The DM's perceived variance is derived as

$$E[\omega^{-1}|\, m_t, y_t] = \frac{\beta + \frac{1}{2}\sum_{\tau=1}^{t} (y_\tau - \mu)^2\ \bar{u}_{t,\tau}}{\alpha + \frac{1}{2}\frac{1-\lambda^t}{1-\lambda} - 1} \tag{B.9}$$

from equations (B.8). As before, we define the weight on the variation realized at time τ as

$$s_\tau = \frac{\frac{1}{2}\,\bar{u}_{t,\tau}}{\alpha + \frac{1}{2}\frac{1-\lambda^t}{1-\lambda} - 1}, \quad \tau \leq t. \tag{B.10}$$

Note that the weights on the past realizations are stochastic due to the compounded idiosyncratic noise $\bar{u}_{t,\tau}$. Furthermore, we can see that $s_\tau = \bar{u}_{t,\tau}\, s_t$ where s_t is the weight on the most recent variation. Thus, it is straightforward to see that

$$E\left[s_\tau\right] = \lambda^{t-\tau}\, s_t, \quad \tau \leq t \tag{B.11}$$

where the expectation is taken over possible realizations of $\bar{u}_{t,\tau}$. This implies that weights on past variations exponentially decay on average.

We can also see that the weight on the most recent observation s_t does not converge to zero as $t \to \infty$. Instead, the limit converges to a positive number tied to the extent of noisy memory. This means that recent surprises keep affecting the DM's perception of ω even after infinitely long learning opportunities. An extreme case of "no memory" clarifies intuition. Suppose the DM's beliefs are formed only based on the current observations despite having observed past variations of y_t. In this case, the DM's beliefs will markedly differ from those of a statistician, although they have observed the same data. Since past data does not affect the DM's beliefs ($\lambda = 0$), it must be that $\alpha_t = \frac{1}{2}$ and $\beta_t = \frac{1}{2}\,(y_t - \mu)^2$ at any t. Therefore, we can see that $E[\omega^{-1}|\, m_t, y_t]$ and $V[\omega^{-1}|\, m_t, y_t]$ keep fluctuating even in the long run. Instead of beliefs converging to a number, they inherit the randomness of the realized y_t.

2.3 Illustrating the Implications of Noisy Memory

In this section, we discuss a numerical example to emphasize the model predictions. We consider the case where y_t is drawn from $\mathcal{N}(0,\, 1)$. We suppose that DM's initial prior belief about ω^{-1} is characterized by $E[\omega^{-1}] = 1$ and $V[\omega^{-1}] = 0.5$, which implies $\alpha = 4$ and $\beta = 3$.

Figure 2.1 shows that recent experiences have bigger influences on the DM's perceived volatility when subject to noisy memory. The top figure describes the average weight put on variations realized in the past when the DM has 20 periods of observations. The x axis is the lag since $t = 20$, so the number 0 and 20 correspond to the most recent and the oldest time periods. The perfect memory case is described by the purple line with a marker. As expected from (B.5), old and new variations all receive the same weights. In inferring the variance of y_t, what matters is how big the extent of surprises are, not when they occurred. In comparison, this prediction does not hold under noisy memory, as captured by the solid blue and dashed green lines. More-recent observations receive higher weights compared to the perfect memory case. Furthermore, weights on older observations decay exponentially, as described by (B.10). This implies that the sequence of observations starts to matter if the DM's perception is based on noisy memory. In the bottom figure, we compare the perceived volatility, $E[\omega^{-1}|\, m_t, y_t]$, under two scenarios. In the first scenario, we suppose that $y_\tau = \mu + \omega^{-\frac{1}{2}}$ for the first 10 periods, followed by $y_\tau = \mu$ in the remaining periods. In the second scenario, we reverse the order. We call these two sequences "Surprise Earlier" and "Surprise Later." Note that if perception is based on perfect memory, the two sequences result in the same level of risk perception at $t = 20$. This is because the order in which surprises are realized should not matter. In comparison, when perception is based on noisy memory, the DM perceives the level of volatility is higher when large variations are recently realized. This prediction is a direct consequence of the decaying weights shown in the top figure.

Figure 2.2 illustrates the DM's uncertainty about the variance of y_t. For both figures,

we report the *expected* level of the posterior uncertainty, averaging over the possible values of $V[\omega^{-1}|\, m_t, y_t]$. As discussed earlier, the posterior uncertainty fluctuates in tandem with the realized variation in y_t. The top figure shows how the posterior uncertainty about ω^{-1} evolves from $t = 0$. Despite observing the same data realizations, the DM's uncertainty about ω^{-1} varies depending on the extent of noisy memory λ. The purple bottom line (labeled with $\lambda = 1.00$) is the perfect memory benchmark. As expected, as learning opportunities accumulate, the DM's uncertainty about ω^{-1} decreases over time and eventually reaches zero. Thus, the DM can learn the actual value of ω through learning. The other lines depict the noisy memory case, with varying degrees of noise in memory. The top red line (labeled with $\lambda = 0.00$) is the no-memory case discussed earlier. Knowledge from past observations is completely lost between periods, so the DM is always as uncertain about ω^{-1} as they are in the initial learning period. We can see that learning is slower under "noisier" memory. The bottom figure makes this point as well. We report the average values of the posterior uncertainty after a long learning period for each level of noisy memory on the x-axis. Noisier memory leads to higher uncertainty about the variance of y_t.

Figure 2.3 displays the impulse response function. We report how the posterior mean of ω^{-1}—the DM's perceived volatility—responds *in the long run* to a positive one standard deviation shock. Different lines correspond to varying degrees of noisy memory λ. The purple line (with a circle marker) is the perfect memory benchmark. As discussed earlier, after a long enough learning period, the DM eventually learns about ω. This means that the DM's perception of ω^{-1} does not change in response to a fluctuation in the realized y_t. This prediction does not hold when $\lambda < 1$. For all the rest of the lines, we see that the DM perceives that y_t is more variable after observing an unusually high realization of y_t. The degree of on-impact response and the dynamics afterward vary depending on the noise in memory. For example, in the no-memory case (the red dotted line), we see the largest on-impact response that immediately dies out the following period. Since the DM is hugely uncertain about the value of ω, a large variation of y_t signals that the variance of

y_t is high. This inference, however, is not carried over to the following periods since her future beliefs do not put weight on past data. The posterior mean is serially correlated in the intermediate case of $\lambda \in (0, 1)$. A one-time experience of high y_t keeps influencing the DM's perception of ω^{-1}.

2.4 Testable predictions

The main prediction of the model we would like to test is captured by the relationship between perceived risk and recent surprises.

$$\beta \equiv \frac{Cov\left[PerceivedRisk_t, RecentSurprises_t\right]}{V[RecentSurprises_t]} \tag{B.12}$$

In our model, perceived volatility is defined as $E[\omega^{-1}|\, m_t, y_t]$, and recent surprises are defined as $(y_t - \mu)^2$. From (B.4) and (B.9), we can immediately see that $\beta = s_t$, where s_t is the weight on the most recent surprises as defined in (B.5) and (B.10).

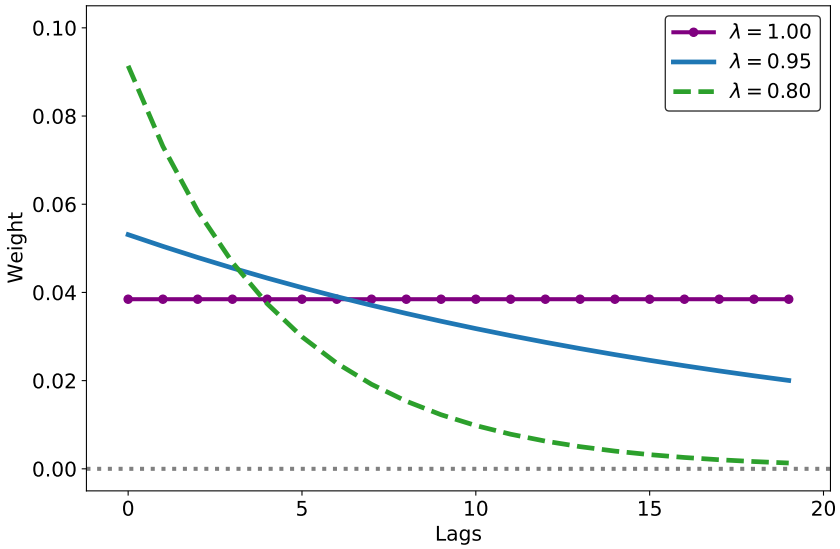
Importantly, the long-run properties of s_t allow us to test if our noisy memory theory is validated in the data. We have shown that β must be zero if DM is learning about the variance with perfect memory. In other words, perceived risk does not fluctuate with recent surprises. This is because the DM comes to have accurate knowledge of the actual variance. In contrast, under noisy memory, it must be that

$$\beta = \left(2\left(\alpha - 1\right) + \frac{1}{1-\lambda}\right)^{-1} \tag{B.13}$$

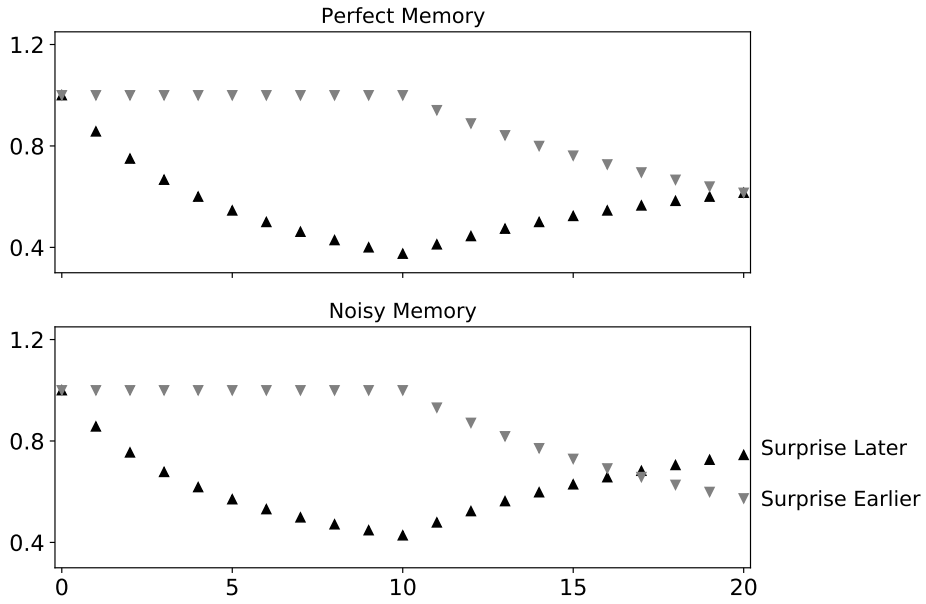
in the long run. Thus, s_t is positive and more positive for noisier memory. With noisy memory, recent surprises continue influencing the DM's view about volatility. Thus, following a larger magnitude of surprises, the DM perceives that extreme realizations are more likely in the future. When the realized surprise is modest, the perceived volatility is lower under noisy memory compared to the perfect memory case. The slope of this relationship is steeper with noisier memory.

Figure 2.1: More Weight on Recent Surprises

(a) Weights on Past Surprises



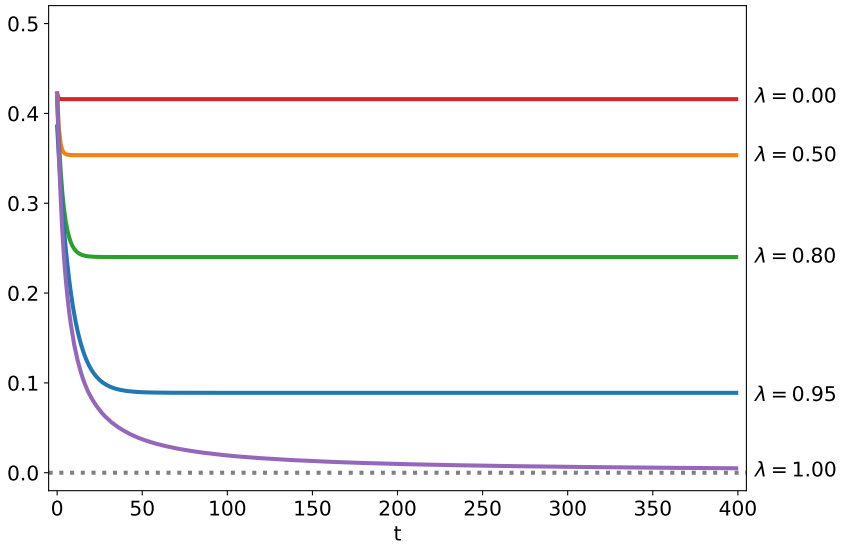
(b) Risk Perception and Order Effects



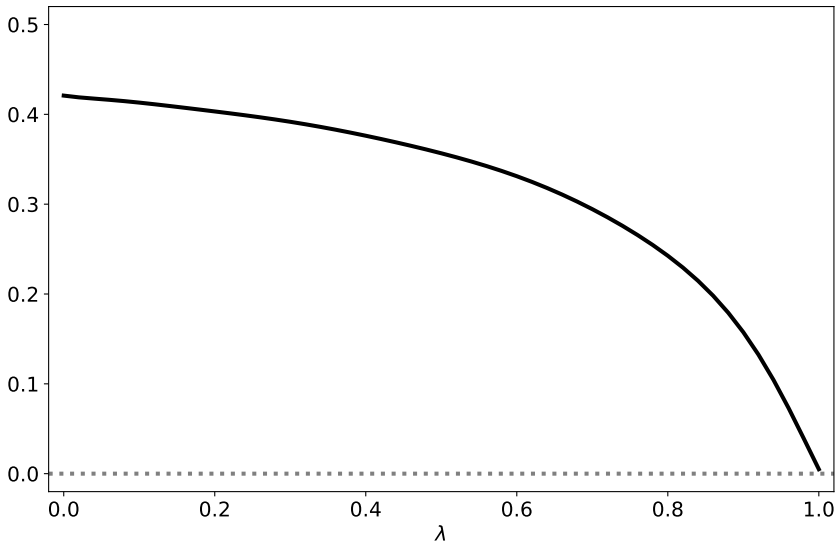
The top figure shows the weights on past surprises. The figure assumes that DM has 20 periods of learning opportunities. Each line corresponds to a different degree of noisy memory λ. The bottom figure reports $E[\omega^{-1}\big|\, m_t, y_t]$ when DM has observed the following sequences. The first scenario, labeled as "Surprise Earlier", assumes $y_\tau = \mu + \omega^{-\frac{1}{2}}$ for $\tau \leq 10$ and $y_\tau = \mu$ for $\tau > 20$. The second scenario, labeled as "Surprise Later", assumes $y_\tau = \mu$ for $\tau \leq 10$ and $y_\tau = \mu + \omega^{-\frac{1}{2}}$ for $\tau > 20$.

Figure 2.2: Posterior Uncertainty about ω^{-1}

(a) Evolution of $V\left[\omega^{-1}|m_t, y_t\right]$

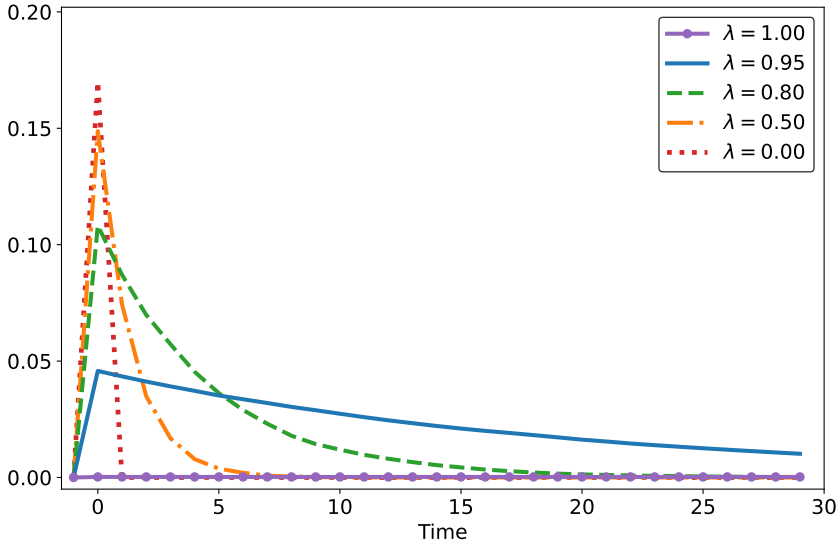


(b) Steady-state values of $V\left[\omega^{-1}|m_t, y_t\right]$



Above figures display the model-predicted posterior uncertainty of ω^{-1}. The top panel shows how the posterior uncertainty evolves for varying degrees of noisy memory λ. The bottom panel shows the steady-state levels of posterior uncertainty for each λ on the x axis.

Figure 2.3: Impulse Response Function of $E\left[\omega^{-1}|m_t, y_t\right]$



The figure displays how the posterior mean of ω^{-1} responds to a one standard deviation increase in the realization of y_t. Each line corresponds to a different degree of noisy memory λ. The figure assumes that DM has had infinite learning periods.

3 Empirical Evidence

We now turn to an empirical evaluation of the model's implications.

3.1 Data

Our data on inflation forecasts come from U.S. Survey of Professional Forecasters (SPF), administered quarterly by the Philadelphia Fed. For a subset of variables, the SPF asks forecasts to submit both point forecasts and a probability density function over realizations of those variables at particular horizons. Probability assessments are available for inflation based on the gross domestic product (GDP) deflator, the core personal-consumption expenditures price index, and the core consumer price index. The latter two series only begin in 2007, so we focus on expectations of inflation based on the GDP deflator, expectations of which have been collected since 1969.

We also collect data on real-time estimates of inflation from the Philadelphia Fed. The SPF is administered shortly after the "advance" release of GDP (including its componenets and price index). The release for quarter $t-1$ occurs late in the first month of quarter $t-1$. We denote the first-release of annualized inflation for quarter $t-1$ as $\pi_{t-1|t}$.[3] The Survey's administrators make this information available to forecasters when taking the survey at time t—it is, thus, the most recently available data on inflation.

Participants in quarter t submit a probability distribution over the level of inflation that will prevail at time τ. Specifically, participants are given B bins, indexed by b, that cover a finite range of the real number line. Each bin has a maximum u_b and minimum ℓ_b value.[4] The probability assigned by participant i that annual inflation at time τ will fall in bin b is

[3] We construct this from real-time estimates of the prices index, P. Therefore, $\pi_{t-1|t} = \left(\left(\frac{P_{t-1|t}}{P_{t-2|t}}\right)^4 - 1\right) \times 100$. We use this same construction when working with forecasts of the price level.

[4] This range is top- and bottom-coded. For the bottom-coded and top-coded bins, we set $u_b = \ell_b$. Thus, our estimate of the variance is likely downward biased. We are exploring alternative approaches in ongoing work.

given by $\pi_{i,t,\tau,b}$. We denote the midpoint of the bin by $m_b = 0.5 \times (u_b + \ell_b)$ and define

$$\mathbb{M}_{i,t,\tau} = \sum_{b=1}^{B} \pi_{i,t,\tau,b} m_b,$$

and their subjective inflaiton uncertainty by

$$\mathbb{V}_{i,t,\tau} = \sum_{b=1}^{B} \pi_{i,t,\tau,b} \left(m_b - \mathbb{M}_{i,t,\tau}\right)^2.$$

This is our primary measure of inflation uncertainty. In practice, τ is either the end of year t, or the following year. Our primary measure of the inflation surprise is given by

$$x_{i,t} \equiv \pi_{t-1|t} - \mathbb{E}_{t-1}\left[\pi_{t-1}\right],$$

In words, this is the surprise component of (annualized) quarterly inflation from quarter $t-2$ to quarter $t-1$. We construct the term $\mathbb{E}_{t-1}[\pi_{t-1}]$ from the point forecasts of the price level, as this is the point forecast solicited by the survey.

Our baseline sample runs from 1992 through 2019. We start in 1992 because the definition of the exact variable being forecast changed. Before 1992, the SPF asked participants about gross national product. Starting in 1992, participants were asked about GDP. We consider an earlier start date (the earliest possible—1969—in a robustness exercise). Starting the sample after the extreme volatility in inflation in the late 1970s/early 1980s also ensures that we are learning about the behavior of forecasts when inflation is relatively stable. We end our analysis before the COVID pandemic for a practical reason—the increase in inflation in 2021 caused many forecasters to put high probability mass in the highest bin. This mechanically makes uncertainty low, but is in fact the result of unchanging bins in the survey.

Table 2.1 shows basic summary statistics of the variables we have constructed. Surprises are close to zero on average, but with a fairly wide standard deviation and, thus,

average absolute size. The average squared forecast revision in our baseline sample is about one and a half percentage points (in annualized inflation terms). Surprises and uncertainty tend to be larger in the earlier part of the sample, consistent with the very volatile underlying processes in the late 70s. Uncertainty deceases later in the calendar year in both samples—hence the quarter-of-year fixed effects—though this is not a significant feature of the size of forecast revisions, since these are not calendar-time forecasts. Uncertainty about the current calendar year is also lower than uncertainty about the following year.

3.2 Estimates

To assess the relationship between inflation surprises and perceived inflation risk, we estimate the following specification

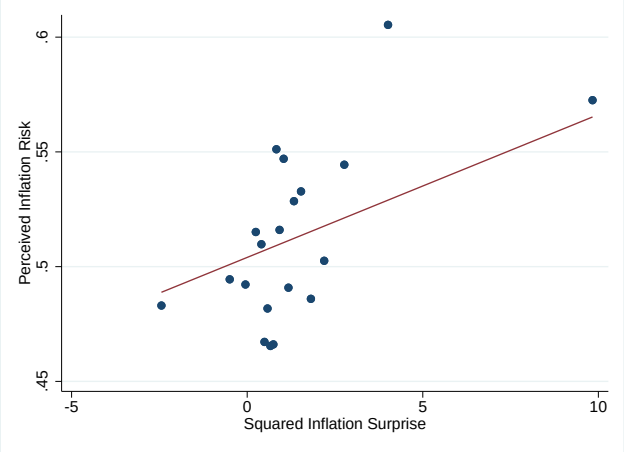
$$\mathbb{V}_{i,t,\tau} = \beta x_{i,t}^2 + \delta_i + \delta_{q(t),\tau} + \varepsilon_{i,t}. \tag{C.14}$$

The term δ_i denotes a forecaster fixed effect, while $\delta_{q(t),\tau}$ is a forecast-horizon $\times$ quarter-of-year fixed effect. We include the latter in order to address the fact that SPF participants are forecasting calendar-time outcomes, so $\tau - t$ becomes smaller later in the quarter, and can thus reduce uncertainty mechanically. In robustness exercises, we show that only focusing on next-year forecasts yields similar results. The coefficient of interest is β—the coefficient on the squared value of the inflation surprise. We scale this term by its unconditional standard deviation in the 1969–2019 sample for ease of interpretation. The theory of section 2 suggests that $\beta > 0$.

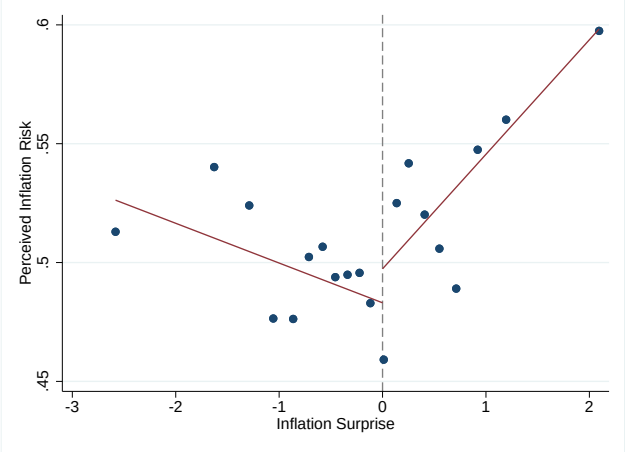
Table 2.2 shows the results. In the first column, we show our baseline estimates during the 1992–2019 sample, using all forecast horizons. The relationship between inflation surprises and perceived inflation risk is positive and statistically significant. In terms of magnitudes, a 1 standard deviation squared inflation surprise leads to an increase in the perceived variance of inflation of 0.05 percentage point. This is roughly ten percent of the

Figure 2.4: Scatter Plot of Baseline Regression

(a) Squared Surprises



(b) Surprises



NOTE. Panel (a) shows a binned scatter plot of equation(C.14) (i.e., all variables have been orthogonalized to the controls in that regression). Panel (b) allows positive and negative revisions to enter with different coefficients. See the note to table 2.2 for more details.

average value of the perceived variance.

The additional columns of table 2.2 confirm that these findings are robust to the alternative specifications considered. The second and third columns split the original sample by forecast horizon—either the current calendar year, or the year after. The results are attenuated somewhat for the current calendar year, though it is worth noting that perceived inflation risk in the current calendar year is much lower for the current calendar year, as seen in table 2.1. The results using the extended sample are similar to the baseline sample, though are also somewhat attenuated.

Figure 2.4 shows a scatter plot of the data underlying our baseline estimates. Both panels in the figure show inflation surprises and inflation risk orthogonalized to the same controls as in our baseline. Panel (a) shows squared surprises, as in our baseline estimates of equation (C.14). Because squaring surprises can skew the importance of larger surprises (e.g. the point on the far right of panel (a)), in appendix 5.2 we show that this relationship continues to hold when using the absolute value of forecast revisions, and when estimating "Huber-robust" regressions, as in the work of Coibion et al. (2018). Panel

(b) shows the non-squared surprises, and allows the coefficients on negative and positive values to be equal, in contrast to our baseline estimates. Despite this relaxation, the same pattern emerges—larger forecast revisions of both signs are associated with larger inflation uncertainty.

4 Conclusion

We propose a theory of how risk perception is formed when subject to noisy memory. We show that perceived risk fluctuates even when the underlying true risk does not change at all. Moreover, we show that recent surprises have an outsized influence on one's perceived risk, while older surprises have an exponentially decreasing influence.

Our theory provides a new implication for the asset pricing literature. Perceived risk has been argued to be important in explaining the historical equity premium and risk-free return. In this literature, the conventional approach is to consider a shock to the variance of the underlying process. For example, Weitzman (2007) argues that such variance shock can generate fluctuating market-wide risk perception, which explains several "puzzles" in the literature. While plausible, this idea did not have a quantitative success. As Bakshi and Skoulakis (2010) argues, explaining features of the asset prices with the variance shock still requires an implausible level of stochasticity that the data does not justify. Our model provides another reason why risk perception fluctuates, and this reason is not tied to the data generating process of the underlying variance.

Table 2.1: Summary Statistics

Baseline Sample (1992–2019)

	Observations	Mean	Std. Dev.	25th Pctile.	75th Pctile.
$s_{i,t}$	6349	-0.18	1.16	-0.86	0.52
$s_{i,t}^2$	6349	1.38	3.14	0.12	1.42
$s_{i,t}^2$, Q1 or Q2	3149	1.43	2.83	0.11	1.48
$s_{i,t}^2$, Q3 or Q4	3200	1.33	3.41	0.13	1.38
$\mathbb{V}_{i,t,\tau}$	6349	0.51	0.55	0.20	0.65
$\mathbb{V}_{i,t,\tau}$, Q1 or Q2	3149	0.57	0.55	0.24	0.73
$\mathbb{V}_{i,t,\tau}$, Q3 or Q4	3200	0.46	0.54	0.16	0.56
$\mathbb{V}_{i,t,\tau}$, current year	3191	0.41	0.47	0.14	0.52
$\mathbb{V}_{i,t,\tau}$, next year	3158	0.61	0.60	0.25	0.77

Extended Sample Sample (1969–2019)

	Observations	Mean	Std. Dev.	25th Pctile.	75th Pctile.
$s_{i,t}$	9363	-0.09	1.52	-0.92	0.71
$s_{i,t}^2$	9363	2.31	8.27	0.16	2.05
$s_{i,t}^2$, Q1 or Q2	4709	2.44	6.65	0.15	2.36
$s_{i,t}^2$, Q3 or Q4	4654	2.18	9.63	0.16	1.97
$\mathbb{V}_{i,t,\tau}$	9363	0.65	0.75	0.21	0.81
$\mathbb{V}_{i,t,\tau}$, Q1 or Q2	4709	0.72	0.79	0.25	0.87
$\mathbb{V}_{i,t,\tau}$, Q3 or Q4	4654	0.58	0.69	0.19	0.75
$\mathbb{V}_{i,t,\tau}$, current year	5486	0.56	0.68	0.19	0.70
$\mathbb{V}_{i,t,\tau}$, next year	3877	0.77	0.81	0.26	0.95

Table 2.2: Empirical Estimations: Inflation Surprises and Perceived Inflation Risk

s_{it}^2	0.0514	0.0431	0.0598	0.0322
	(0.0218)	(0.0204)	(0.0262)	(0.0110)
Observations	6349	3179	3145	9363
R^2	0.531	0.500	0.581	0.474
Sample	1992–2019	1992–2019	1992–2019	1969–2019
Horizons	0 & 1	0	1	0 & 1

NOTE. This table shows estimates of equation (C.14). The left-hand side is expressed in percentage points, while $x_{i,t}^2$ has been scaled by its unconditional standard deviation (which can be seen in the second panel of table 2.1). The first column shows our baseline estimates. The second and third columns show our results for forecasts of inflation in the year the forecast is made, and in the subsequent year, respectively. The final column includes all observations since 1969. Standard errors clustered by i and t are shown in parentheses. All regressions include the fixed effects mentioned in the description of equation (C.14).

5 Appendix

5.1 Detailed Derivations

We derive how beliefs about ω (the inverse of the variability of y_t) evolve if the "memory process" is formed according to (B.6). We show this recursively.

At $t = 1$, DM starts with a prior belief denoted as $\omega \sim \Gamma(\alpha,\, \beta)$. After observing y_1, the posterior belief is formed as

$$\omega|\, y_1 \sim \Gamma\left(\alpha + \underbrace{\frac{1}{2}}_{\equiv\alpha_1},\, \beta + \underbrace{\frac{1}{2}(y_1 - \mu)^2}_{\equiv\beta_1}\right)$$

from which we can see that $\beta_1|\, \omega \sim \Gamma(\alpha_1,\, \omega)$. The memory variable is defined as

$$m_2 = \beta_1\, u_2$$

where $u_2 \sim \mathcal{B}(\lambda\alpha_1,\, (1-\lambda)\,\alpha_1)$. Then, it must be that the likelihood function of m_2 is derived as $m_2|\, \omega \sim \Gamma(\lambda\,\alpha_1,\, \omega)$.

Given this structure, we can derive the prior belief at $t = 2$ as follows.

$$\begin{aligned}\omega|\, m_2 &\propto f(m_2|\, \omega)\; f(\omega) \propto \left(\omega^{\lambda\,\alpha_1}\, \exp^{-\omega\, m_2}\right)\left(\omega^{\alpha-1}\, \exp^{-\beta\,\omega}\right) \\ &\propto \Gamma\left(\alpha + \underbrace{\lambda\,\alpha_1}_{\equiv\alpha_2^m},\, \beta + \underbrace{m_2}_{=\beta_1\, u_2 \equiv \beta_2^m}\right)\end{aligned}$$

Thus, the memory variable m_2 maintains the prior distribution to be a Gamma distribution. I use the notation $\omega|\, m_2 \sim \Gamma(\alpha + \alpha_2^m,\, \beta + \beta_2^m)$ to denote this distribution. After observing

y_2, DM updates her beliefs, which again are described as a Gamma distribution.

$$\omega|\, m_2, y_2 \sim \Gamma\left(\alpha + \underbrace{\alpha_2^m + \frac{1}{2}}_{\equiv\alpha_2},\, \beta + \underbrace{\beta_2^m + \frac{1}{2}\left(y_2 - \mu\right)^2}_{\equiv\beta_2}\right)$$

From $\beta_2^m|\,\omega \sim \Gamma\left(\alpha_2^m, \omega\right)$ and $\frac{1}{2}\left(y_2-\mu\right)^2\big|\,\omega \sim \Gamma\left(\frac{1}{2}, \omega\right)$, we get $\beta_2|\,\omega \sim \Gamma\left(\alpha_2, \omega\right)$. The memory variable is defined as

$$m_3 = \beta_2\, u_3$$

where $u_3 \sim \mathcal{B}\left(\lambda\alpha_2,\, (1-\lambda)\,\alpha_2\right)$. Then again, we see that

$$\omega|\, m_3 \sim \Gamma\left(\alpha + \underbrace{\lambda\,\alpha_2}_{\equiv\alpha_3^m},\, \beta + \underbrace{\beta_2\, u_3}_{\equiv\beta_3^m}\right)$$

Recursively, we can then express any time t prior as

$$\omega|\, m_t \sim \Gamma\left(\alpha + \alpha_t^m,\, \beta + \beta_t^m\right)$$

where $\beta_t^m|\,\omega \sim \Gamma\left(\alpha_t^m, \omega\right)$ is held, and we can also express the posterior as

$$\omega|\, m_t, y_t \sim \Gamma\left(\alpha + \alpha_t,\, \beta + \beta_t\right)$$

where

$$\begin{aligned}\alpha_t &= \alpha_t^m + \frac{1}{2}\\ \beta_t &= \beta_t^m + \frac{1}{2}\left(y_t - \mu\right)^2\end{aligned}$$

Then, using the distribution of $\beta_t^m|\,\omega$, we see that $\beta_t|\,\omega \sim \Gamma\left(\alpha_t, \omega\right)$. Using the definition of

the memory variable m_{t+1} in (B.6), we can also derive that $m_{t+1}|\,\omega \sim \Gamma\left(\lambda\,\alpha_t,\,\omega\right)$. Finally, we derive $\omega|\,m_{t+1}$ below.

$$\begin{aligned}\omega|\,m_{t+1} &\propto f\left(m_{t+1}|\,\mu,\omega\right)\,f\left(\omega|\,\mu\right)\\ &\propto \left(\omega^{\lambda\,\alpha_t}\,\exp^{-\omega\,m_{t+1}}\right)\left(\omega^{\alpha-1}\,\exp^{-\beta\,\omega}\right) \propto \Gamma\left(\alpha+\alpha^m_{t+1},\,\beta+\beta^m_{t+1}\right)\end{aligned}$$

where

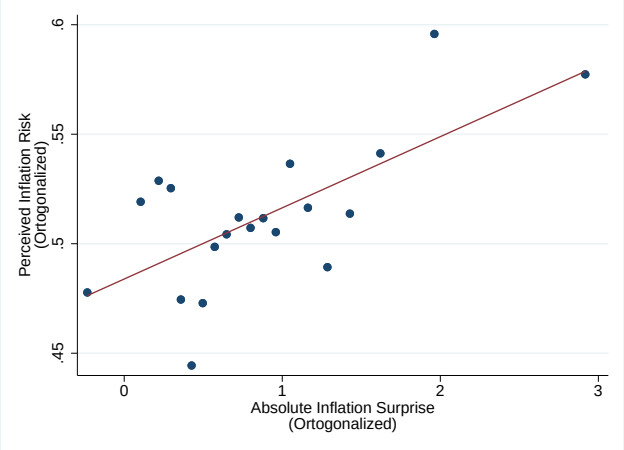
$$\begin{aligned}\alpha^m_{t+1} &= \lambda\,\alpha_t\\ \beta^m_{t+1} &= m_{t+1} = \beta_t\,u_{t+1}\end{aligned}$$

5.2 Robustness Checks

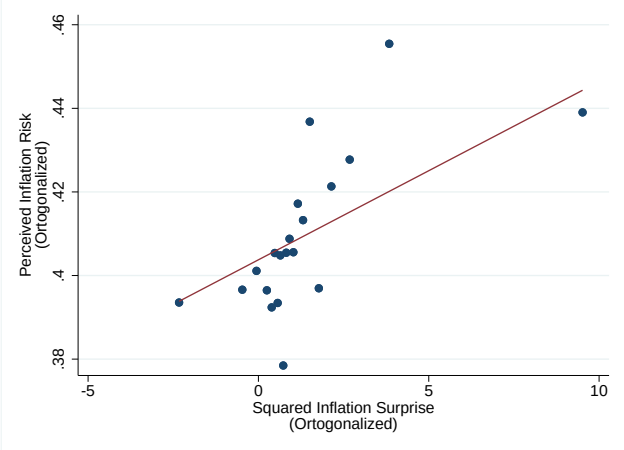
In figure 2.5, we show alternative versions of figure 2.4 in which we replace squared surprises with absolute revisions (panel (a)) or use Huber weights (panel (b)). The numerical estimates associated with these figures is presented in table 2.3

Figure 2.5: Scatter Plot of Baseline Regression: Robustness

(a) Absolute Surprises



(b) Huber-Robust



NOTE. Panel (a) shows a binned scatter plot of equation(C.14) (i.e., all variables have been orthogonalized to the controls in that regression) except that we have replaced the squared surprise with the absolute value of the surprise. Panel (b) shows estimates of our baseline specification using Huber weights, following Coibion et al. (2018).

Table 2.3: Empirical Estimation: Robusntess

x_{it}^2	0.0514	0.0352	
	(0.0218)	(0.0129)	
$\lvert x_{i,t}\rvert$			0.0346
			(0.0136)
Observations	6349	6136	6349
R^2	0.531	0.740	0.531
Sample	1992–2019	1992–2019	1992–2019
Specification	Baseline	Huber	Abs. Rev.

NOTE. This table shows estimates of equation (C.14). The first column is a replication of the first column (our baseline) of table 2.2—see the note to that table for more details. In the second column, we use Huber weights to eliminate the role of outliers. In the third column, we replace squared surprises with absolute surprises.

Chapter 3: Macroeconomic Expectations and Cognitive Noise

1 Introduction

Despite being commonly held, the empirical validity of the full-information rational expectations (FIRE) assumption has long been questioned. According to FIRE, forecasts made by economic agents should not include predictable errors because all information is used efficiently. However, an expanding empirical literature has found that this prediction does not hold, even for professional forecasters who presumably have ample information and advanced tools.

In particular, recent studies argue that econometricians can predict errors that forecasters will make based on the latter's recent forecast revisions. However, a puzzle emerges. Coibion and Gorodnichenko (2015) finds that the average forecast tends to *undershoot* realizations when forecasters revise their projections upward. In comparison, Bordalo, Gennaioli, Ma, and Shleifer (2020b) find that when an individual forecaster revises her projection upward, her forecast tends to *overshoot* realizations. Of course, the revision of the average forecast is more muted than the revision of the individual one, as the average forecast averages different views. However, it is not apparent how to explain the flip in the direction of predictability.

This paper shows that noisy information can account for the seemingly contradictory pattern. While forecasters have access to a vast amount of information, they have finite capacity to process it. I propose that forecasts are based on the *mental representation* of available information, not all available information. The mental representation can be considered a noisy summary of information; it is composed of both the mind's distilled understanding of available information and the noise uncorrelated with the distillation.

This process is endogenously determined to maximize forecast accuracy given processing constraints.

To explain these patterns in both average and individual forecasts, I distinguish between two types of information: external and internal. External information can be looked up; it includes data releases, news articles, press conferences, and the like. Internal information is stored in forecasters' memory; it consists in their accumulated knowledge from past forecasting experience. Importantly, both types of information are noisy: forecasts are formed based on the mental representation of available external and internal information. Using this model, I estimate the extent of information frictions using professional forecasters' projections of the overall US economy. Furthermore, I explore the monetary-policy implications of the estimated model.

Conventional models of information frictions assume that only external information is noisy. These models can explain why consensus forecasts undershoot. When facing a new set of external information, forecasters, on average, are less responsive than under FIRE because the new information is noisy. At the individual level, forecasters make projections efficiently given the noise. Therefore, such revisions do not predict systematic errors in individual forecasts.

I show that by adding noisy internal information, I can explain the predictability of both average and individual forecasts. Noisy external information generates the consensus-level pattern, as in Coibion and Gorodnichenko (2015), while noisy internal information generates the individual-level pattern. When forecasters cannot freely access their internal information, their prior knowledge resolves less uncertainty about the forecast variable. Thus, forecasters put extra weight on new information. This extra sensitivity to new information explains why individual forecasts tend to overshoot more than conventional information-friction models predict.

Jointly considering both types of noisy information is crucial for understanding the extent of information frictions. This is because the sensitivity with which forecasts are

revised depends on both types of information. Forecast revisions can be sensitive for two reasons: external information is not very noisy, or internal information is quite noisy. I show that the extent of the two types of noise determines the revision pattern of average and individual forecasts. Not considering both types of noisy information would lead one to misestimate the degree of information frictions.

A direct implication is that Coibion and Gorodnichenko (2015) underestimates the extent of information frictions. The authors argue that the severity of information constraints can be inferred from the revision pattern of consensus forecasts. However, since this methodology implicitly assumes that accessing internal information is costless, it does not account for the extra weight on new information arising from noisy internal information. Therefore, the methodology proposed by Coibion and Gorodnichenko (2015) misattributes this extra sensitivity to less severe information frictions.

To improve the model's empirical validity, I extend the model along one more dimension: forecasters learn about the long-run mean. Instead of assuming that forecasters are fully aware of where the forecast variable reverts to, I assume they learn about the long-run steady state over time. While more than one parameter determines the steady state, I focus on the mean because knowledge about the mean is essential for forecasters making long-term forecasts.

In the proposed model, forecasters are perpetually uncertain about the long-run mean. If accessing internal information is costless, forecasters eventually learn about the mean. This is the basis on which many models assume forecasters have perfect awareness of the model parameters. However, as my other work Azeredo da Silveira et al. (2020) shows, forecasters' knowledge about the mean is imperfectly accumulated over time when internal information is not perfectly accessible. In this case, learning persists even after extensive learning opportunities.

Using the extended model, I estimate the degree of information frictions. The model is applied to professional forecasters' projections of US economic variables related to output,

the price level, the labor and housing markets, and borrowing costs. For each macroeconomic variable, the two constraints — one in processing external information and the other in processing internal information — are inferred from the forecast-revision patterns at the consensus and individual levels. I find that the extent of information frictions is more substantial than what the conventional information-frictions literature finds: my estimate of the constraint in processing external information is twice as large as that of Coibion and Gorodnichenko (2015). I also show that the estimated model explains sizable shares of the variation in forecasts and revisions, both in the cross sections and in the time series.

Then, I explore the macroeconomic implications of the proposed information frictions — in particular, the implications for how inflation is determined. I use a standard New Keynesian model in which firms set prices based on their macroeconomic expectations. Using this framework, I show how the inflation process varies with the assumption of the expectation-formation process. Furthermore, I investigate the operation of monetary policy in balancing the trade-off between inflation and output stabilization.

If firms are subject to the costly information proposed in this paper, stabilizing inflation can be more challenging than under FIRE. The key reason is that inflation expectations are unanchored because internal information is costly to process. Since firms do not have perfect awareness of the long-run economy, their beliefs about it fluctuate with persistence. This additional fluctuation is transmitted through their price setting, making aggregate inflation volatile. In this economy, a monetary policy strongly emphasizing inflation stabilization can more effectively guide economic agents' long-run expectations.

In proposing a new expectation-formation model, I provide a parsimonious explanation for the puzzling features of survey forecasts. In the model I present, one type of information friction keeps economic agents from making forecasts consistent with FIRE: finite capacity to process a vast amount of available information. In comparison, previous proposals in the literature resort to a non-Bayesian assumption *in addition to* information frictions to explain the forecast-revision patterns discussed in this paper. For example, rep-

resentative heuristics (Bordalo, Gennaioli, Ma, and Shleifer (2020b)), misspecification of the model (Angeletos, Huo, and Sastry (2021)), and desire to stand out from the crowd (Gemmi and Valchev (2021)) have been proposed. While these may be plausible and insightful proposals, it is unclear how economic agents come to have such biases. Furthermore, I show that the model I present explains features of survey forecasts regarding the forecast horizon that these previous proposals cannot explain.

The findings from this paper also shed light on the formation of *long-run* inflation expectations. It has long been recognized that economic agents' inflation expectations affect the inflation process. Thus, the implementation of monetary policy should carefully consider the exact nature of expectation formation (Orphanides and Williams (2004)). However, recent literature argues that expectations about long-run inflation are crucial to understanding past inflation dynamics (Carvalho, Eusepi, Moench, and Preston (2022), Hazell, Herreño, Nakamura, and Steinsson (2022)) and have important monetary-policy implications (Gàti (2021)). A popular proposal in the literature is that economic agents learn about unobservable stochastic trends from the current economy (K. Crump, Eusepi, Moench, and Preston (2021), Farmer, Nakamura, and Steinsson (2021)). While plausible, this idea predicts that economic agents should have well-anchored long-run expectations in response to a sudden spike in inflation if they have experienced low and stable inflation for a long period. In this paper, seemingly anchored long-run inflation expectations can start moving when agents witness bouts of high inflation. This prediction is consistent with experimental studies documenting fluctuations in long-term beliefs in a stable-trend environment (Afrouzi, Kwon, Landier, Ma, and Thesmar (2020)).

More generally, this paper contributes to our understanding of how cognitive limitations affect economic agents' beliefs and decisions. In various fields of economics, cognitive limitations have been proposed to explain several seemingly unrelated patterns (Woodford (2020)). In macroeconomics, rational-inattention theories have been proposed to explain why macroeconomic variables respond to fluctuations in the economy (for exam-

ple, monetary-policy shocks) with a long delay (Sims (2003), Mackowiak and Wiederholt (2009)). In behavioral economics, cognitive uncertainty has been proposed as a unifying explanation for several patterns often viewed as distinct phenomena (Enke and Graeber (2019)). I contribute to this literature by showing that cognitive limitations help us understand the puzzling patterns of survey forecasts emphasized in the macroeconomic literature on expectations.

The paper proceeds as follows. Section 2 presents a model of expectation formation in which forecasts are based on the mental representation of available information. Section 3 discusses what representation is optimal given the information constraints. Section 4 presents the model prediction of the forecast-revision patterns and the estimation strategy. Section 5 describes an extension of the expectation model. Section 6 presents the structural-estimation results. Section 7 describes the illustrative macroeconomic model and discusses the monetary-policy implications. Section 8 concludes.

2 A Model of Mental Representation

In this section, I introduce a model of mental representation (that is, a noisy summary of available information). I describe how a vast amount of information is processed and stored in memory.

2.1 The Forecasting Problem

Consider macroeconomic variable y_t, which is the sum of persistent and transitory components. I assume that

$$y_t = z_t + \eta_t,$$

where η_t is i.i.d, and

$$z_t = (1-\rho)\ \mu + \rho\, z_{t-1} + \epsilon_t,$$

where μ is the long-run mean of z_t, ρ is the serial correlation of z_t (with $|\rho| < 1$), and ϵ_t is an i.i.d. sequence drawn from Gaussian distribution $\mathcal{N}\,(0,\, \sigma_\epsilon^2)$. I assume y_t is observable but z_t and η_t are not directly observed. I assume that all values of parameters describing the stochastic process are known.

The forecasters' problem is to produce projections for future realizations of y_t. The loss from incorrectly forecasting is described by the expected value of a quadratic loss function:

$$E[\sum_{t=0}^{\infty} \beta^t \sum_{h=1}^{H} (y_{t+h} - F_{i,t}\, y_{t+h})^2] \tag{B.1}$$

Here, $F_{i,t}\, y_{t+h}$ is decision-maker (DM) i's forecast of y_{t+h}. Forecasters make projections up to H periods ahead. The expectation operator E is over every possible piece of information available at time t and is described in the remaining section.

Available information. I categorize information into two types: external and internal. Forecasters can look up external information. It includes quantitative and qualitative information, such as data releases, press conferences, and market reports. Internal information is in forecasters' memory —- that is, their past cognitive state.

Cognitive constraints: mental representation. Forecasters' external and internal information is high dimensional and complex. They have a finite capacity to process such data. To capture this constraint, I introduce the notion of mental representation from psychology and cognitive science; it can be considered a noisy summary of information. I propose that forecasters base their projections on the mental representation of available information instead of all available information. The original complex data is distilled into a simpler

form and compounded with random noise, which makes the representation imprecise. This representation is optimally determined, as discussed below.

2.2 Mental Representation of External Information

External information. The underlying state z_t is partially revealed by many pieces of quantitative and qualitative information. Examples of quantitative information are historical realizations of past y_t or other variables relevant for predicting z_t. Qualitative information includes opinions and market commentaries. All such information that is at least somewhat informative about the value of z_t is stored in a large vector N_t. The relationship between N_t and z_t is described as follows:

$$N_t = R \cdot z_t + \nu_t \tag{B.2}$$

R is a constant vector, and $\nu_t \sim \mathcal{N}(0, V)$ for some positive definite matrix V.

Imprecise representation. DM uses various kinds of information in N_t when making forecasts of y_t. I assume that how precisely DM's forecasts depend on this external information is constrained. In particular, I assume that knowledge from N_t is described as follows:

$$n_{i,t} = K_t \cdot N_t + u_{i,t} \tag{B.3}$$

Here, K_t is a matrix (possibly with many fewer rows than the number of elements in N_t) and $u_{i,t} \sim \mathcal{N}(O, \Sigma_{u,t})$ for some positive semidefinite matrix $\Sigma_{u,t}$. The noise $u_{i,t}$ is not correlated with z_t and is idiosyncratic to each forecaster.

The matrices K_t and $\Sigma_{u,t}$ are endogenously determined subject to a constraint. The degree of precision of the mental representation $n_{i,t}$ is measured with the Shannon mutual

information between $n_{i,t}$ and N_t, denoted as $\mathcal{I}(n_{i,t};\, y_t)$.[1] More inaccurate representation is captured by lower mutual information between the two random variables. I assume that the precision of mental representation is constrained as follows:[2]

$$I(n_{i,t}; N_t) \leq -\frac{1}{2} \ln \phi_n \tag{B.4}$$

Here, $\phi_n \in (0, 1)$ parameterizes the upper bound of the mutual information that is taken as given. One can see that a higher ϕ_n allows lower mutual information, thereby constraining the accuracy of the mental representation.

If $\phi_n \to 0$, then forecasts are accurately based on information in N_t. In this case, K_t is an identity matrix (whose dimension is equivalent to the number of rows in N_t) and $\Sigma_{u,t}$ is a zero matrix (with the same dimension as K_t). With $\phi_n > 0$, forecasts are based on the approximate representation of N_t, as K_t may have many fewer rows than the number of elements in N_t and at least some of the diagonal elements of $\Sigma_{u,t}$ are positive. When $\phi_n \to 1$, forecasts are not based on information in N_t, since the representation is infinitely inaccurate.

2.3 Mental Representation of Internal Information

Internal information. In addition to external information N_t, I assume that DM has access to internal information such as her past cognitive state. I denote the internal information accessible at t as $(m_{i,t-1}, n_{i,t-1})$. As discussed earlier, $n_{i,t-1}$ is the mental representation of the news vector N_{t-1}. Meanwhile, $m_{i,t-1}$ is the knowledge carried through $t-1$ before observing N_{t-1}. One can think of $m_{i,t-1}$ as the memory stock of knowledge, and its

[1] This metric captures how "close" $n_{i,t}$ is to N_t. If $\mathcal{I}(n_{i,t};\, y_t)$ is close to zero, then it means knowing $n_{i,t}$ is not informative about N_t. If, on the other hand, the metric is close to infinity, then information delivered by $n_{i,t}$ about N_t is perfectly accurate.

[2] The proposed cost function is different from what is typically assumed in the rational-inattention literature. There, it is assumed that DM can arrange to receive a signal $n_{i,t}$ at time t, conditioning on all the signals till time $t-1$. That is, the cost is assumed to be proportional to $\mathcal{I}(n_{i,t}; y_t | n_{i,t-1}, \cdots, n_{i,0})$. As will be clear from the rest of the model, I consider an environment in which the past realized values of $n_{i,t}$ are not freely available. Therefore, I assume that external information is processed independently of the cognitive state.

evolution will be discussed shortly.

Imperfect representation. I assume the internal information can be represented as follows:

$$m_{i,t} = \Lambda_t \cdot \begin{pmatrix} m_{i,t-1} \\ \\ n_{i,t-1} \end{pmatrix} + \omega_{i,t} \tag{B.5}$$

Here, Λ_t is a matrix that may have fewer rows than $(m_{i,t-1}, n_{i,t-1})$ and $\omega_{i,t}$ is an i.i.d. sequence that is uncorrelated with $(m_{i,t-1}, n_{i,t-1})$ and drawn from the Gaussian distribution $\mathcal{N}(O, \Sigma_{\omega,t})$ for some positive semidefinite matrix $\Sigma_{\omega,t}$.

The two matrixes Λ_t and $\Sigma_{\omega,t}$ are chosen optimally subject to the constraint. The extent of noise in the mental representation $m_{i,t}$ is measured with the Shannon mutual information between $m_{i,t}$ and $(m_{i,t-1}, n_{i,t-1})$. The lower mutual information captures a more inaccurate representation of internal information. In parallel with (B.4), I assume that the accuracy of the representation is constrained as follows:

$$I(m_{i,t}; m_{i,t-1}, n_{i,t-1}) \leq -\frac{1}{2} \ln \phi_m \tag{B.6}$$

Here, $\phi_m \in (0, 1)$ is taken as given. A higher ϕ_m means a more constrained representation.

If $\phi_m \to 0$, forecasts are accurately based on internal information. The corresponding mental representation is when Λ_t is an identity matrix and $\Sigma_{\omega,t}$ is a zero matrix. With $\phi_m > 0$, forecasts rely on imperfect representation of $(m_{i,t-1}, n_{i,t-1})$. When $\phi_m \to 1$, forecasts are not based on internal information, since the represented information is completely inaccurate.

2.4 Forecasts Based on Represented Information

We have seen how external and internal information is mentally represented. For brevity, I refer to $n_{i,t}$ as *noisy news* (that is, an imperfect representation of external information) and $m_{i,t}$ as *noisy memory* (that is, an imperfect representation of internal information). I consider the representation to be noisier if the accuracy of representation is more constrained (higher ϕ_n or ϕ_m).

Bayesian forecasts subject to information constraints. I assume that forecasts are Bayesian efficient given the noisy news $n_{i,t}$ and noisy memory $m_{i,t}$. That is, while the bottleneck is in processing the complex, high-order information, forecasters have expertise in combining $n_{i,t}$ and $m_{i,t}$. The conditional distribution is derived using the usual Kalman filter formula.

Implications of the linear-Gaussian structure. The linear-Gaussian structure of $n_{i,t}$ and $m_{i,t}$ implies that DM's beliefs about the past and current realizations of z_t take the form of a Gaussian distribution. In other words, $(z_0, \cdots, z_t)\,|m_{i,t}$ and $(z_0, \cdots, z_t)\,|m_{i,t}, n_{i,t}$ are both Gaussian. (The second moment of the Gaussian distribution captures the uncertainty DM feels, which depends on the severity of news noise and memory noise.) Since DM's beliefs about the past and current realizations are Gaussian, DM's belief about future realizations is also Gaussian.

I introduce the following notations to denote DM's beliefs about the state z_τ implied by her cognitive states:

$$z_\tau|\, m_{i,t} \sim \mathcal{N}\left(z^m_{i,i,\tau|t},\, \Sigma^m_{\tau|t}\right)$$
$$z_\tau|\, m_{i,t}, n_{i,t} \sim \mathcal{N}\left(z_{i,i,\tau|t},\, \Sigma_{\tau|t}\right)$$

The top distribution refers to the (beginning of period t) prior belief conditioned on the

memory state at time t. The superscript m indicates that beliefs are based on memory alone. The bottom distribution is the posterior belief after observing $n_{i,t}$ (and is denoted without the superscript m).

Then, the optimal forecasts of y_{t+h} will be

$$F_{i,t}\, y_{t+h} = \left(1 - \rho^h\right) \mu + \rho^h\, z_{i,t|t},$$

from which the mean squared error from forecasting y_{t+h} equals

$$E[(y_{t+h} - F_{i,t}\, y_{t+h})^2] = \rho^{2h}\, \Sigma_{t|t},$$

where the expectation is over the entire joint probability distribution of possible values of z_t, $m_{i,t}$, and $n_{i,t}$. The average losses from inaccurate forecasting are proportional to $\Sigma_{t|t}$. The loss function (B.1) then reduces to

$$\sum_{t=0}^{\infty} \beta^t \left[q \cdot \Sigma_{t|t}\right], \tag{B.7}$$

where $q \equiv \frac{\rho^2(1-\rho 2H)}{1-\rho^2}$ is a constant known to DM.

2.5 The Nature of Information Frictions

In conventional models of information frictions, forecasters have noisy (or dispersed) information about the state of the economy because they observe the state with idiosyncratic errors (Woodford (2003)). The usual interpretation of this assumption is that forecasters have some fragmented information about the state and no one knows the state perfectly. It is typical to assume further that forecasters store their information and access it in any future period.

In contrast, the information friction in this paper is a cognitive constraint. It is not that forecasters have different sources of information per se but that random cognitive noise

enters while processing the vast set of information. Therefore, even with access to the same information, forecasters have a somewhat different understanding or interpretation of the data, as in Sims (2003).

Importantly, a similar cognitive constraint also applies to information stored in forecasters' memory. In the same way that basing forecasts on all available external information is costly, it is mentally costly to base one's forecasts on all available internal information. Given this constraint, their prior knowledge is imperfectly accessed when they make new projections.

3 The Optimal Mental Representation

We have seen that DM bases her forecasts on two types of information: mental representation of internal information ($m_{i,t}$) and mental representation of external information ($n_{i,t}$). In this section, I discuss the optimal structure of $m_{i,t}$ and $n_{i,t}$.

3.1 The Optimization Problem

The cognitive process is described by the sequence of $\{K_t, \Sigma_{u,t}, \Lambda_t, \Sigma_{\omega,t}\}_{t=0}^{\infty}$. The optimal sequence minimizes the loss function (B.7) subject to the information environment (B.3), (B.4), (B.5), and (B.6).

3.2 Optimal Representation of Noisy News

The optimal $n_{i,t}$ is one-dimensional and has the following structure.

Proposition 2. *$\tilde{n}_{i,t}$ is the optimal representation of $N_{i,t}$ such that*

$$\tilde{n}_{i,t} = \kappa_t \cdot E[\, z_t |\, N_t] + \tilde{u}_{i,t} \tag{C.8}$$

for some positive scalar $\kappa_t \in [0, \bar{\kappa}_t]$ and idiosyncratic noise $\tilde{u}_{i,t}$ drawn from $\mathcal{N}\left(0,\, \sigma^2_{u,t}\right)$.

Proof. See Appendix 10.2. □

Intuitively, the optimal representation of $n_{i,t}$ should only capture information in N_t that is useful for predicting z_t. This is because other information in N_t uses up resources but does not further increase the forecast accuracy. Since $z_t|\, N_t$ follows a Gaussian distribution, such information is summarized in the first moment. Therefore, $\tilde{n}_{i,t}$ encodes $E[\, z_t|\, N_t]$, which is denoted as follows without loss of generality:

$$E[\, z_t|\, N_t] = z_t + \tilde{\nu}_t$$

Here, $\tilde{\nu}_t \sim \mathcal{N}\,(0,\, \sigma_\nu^2)$ for some non-negative σ_ν^2 that is taken as given and known to DM.

As one can see from (C.8), there are combinations of κ_t and $\sigma_{u,t}^2$ that imply the same posterior distribution $z_t|\, m_{i,t}, \tilde{n}_{i,t}$ for any given $m_{i,t}$. Therefore, I impose a normalization so that κ_t alone captures the accuracy of the representation. I assume that

$$Cov\,[\, z_t, \tilde{n}_{i,t}|\, m_{i,t}] = V[\, \tilde{n}_{i,t}|\, m_{i,t}],$$

in which case the posterior uncertainty is determined as

$$\Sigma_{t|t} = (1 - \kappa_t)\, \Sigma_{t|t}^m$$

for a given prior uncertainty $\Sigma_{t|t}^m$. That is, observing $\tilde{n}_{i,t}$ reduces the uncertainty about z_t by a factor of $1 - \kappa_t$. This normalization pins down $\sigma_{u,t}^2$ as the following function of κ_t:

$$\sigma_{u,t}^2 = \kappa_t\, (1 - \kappa_t)\, \Sigma_{t|t}^m - \kappa_t^2\, \sigma_\nu^2$$

One can then see that any $\kappa_t \in \left[0,\, \frac{\Sigma_{t|t}^m}{\Sigma_{t|t}^m + \sigma_\nu^2}\right]$ ensures that the resulting $\sigma_{u,t}^2$ is non-negative.

The value of κ_t is determined by the accuracy constraint (B.4). Given the optimal structure of $n_{i,t}$, the mutual information between $n_{i,t}$ and N_t equals $\mathcal{I}\,(\tilde{n}_{i,t};\, z_t + \tilde{\nu}_t)$. Then,

we can pin down κ_t as a function of ϕ_n:

$$\kappa_t = \frac{\Sigma^m_{t|t}}{\Sigma^m_{t|t} + \frac{\phi_n}{1-\phi_n}\left(V[z_t] + \sigma^2_\nu\right) + \sigma^2_\nu} \tag{C.9}$$

We can see that noisier news implies lower κ_t and higher posterior uncertainty. Also note that after long enough learning, the subjective uncertainty $\Sigma^m_{t|t}$ and $\Sigma_{t|t}$ converge to a positive steady-state level for all t. Accordingly, $\kappa_t \to \kappa$.

3.3 Optimal Representation of Noisy Memory

The optimal $m_{i,t}$ is one-dimensional and has the following structure.

Proposition 3. *$\tilde{m}_{i,t}$ is the optimal representation of $(m_{i,t-1}, n_{i,t-1})$ such that*

$$\tilde{m}_{i,t} = \lambda_t \cdot z_{i,i,t|t-1} + \tilde{\omega}_{i,t+1} \tag{C.10}$$

for some positive scalar $\lambda_t \in [0,1]$ and idiosyncratic noise $\tilde{\omega}_{i,t}$ drawn from $\mathcal{N}\left(0,\, \sigma^2_{\omega,t}\right)$.

Proof. See Appendix 10.2. □

The intuition for deriving the optimal structure is similar to the derivation of $\tilde{n}_{i,t}$. The optimal representation of $m_{i,t}$ captures information in $(m_{i,t-1}, n_{i,t-1})$ that is useful for predicting z_t. Since $z_t |\, m_{i,t-1}, n_{i,t-1}$ follows a Gaussian distribution, such information is summarized in the first moment. Therefore, $\tilde{m}_{i,t}$ encodes $E[\, z_t |\, m_{i,t-1}, n_{i,t-1}]$, which is expressed as $z_{i,t|t-1}$.

As one can see from (C.10), there are combinations of λ_t and $\sigma^2_{\omega,t}$ that imply the same prior distribution $z_t |\, \tilde{m}_{i,t}$. Therefore, I impose a similar type of normalization assumption as I did for noisy news so that the accuracy of the representation is captured by λ_t alone. I impose the restriction that

$$Cov\left[z_t, \tilde{m}_{i,t}\right] = V[\tilde{m}_{i,t}],$$

in which case $V[z_{i,i,t|t-1}\big|\,\tilde{m}_{i,t}] = (1-\lambda_t)\,V[z_{i,i,t|t-1}]$. That is, observing $\tilde{m}_{i,t}$ reduces the uncertainty about $z_{i,i,t|t-1}$ by a factor of $1-\lambda_t$. This pins down $\sigma^2_{\omega,t}$ as a function of λ_t in the following form:

$$\sigma^2_{\omega,t} = \lambda_t\,(1-\lambda_t)\,V[z_{i,i,t|t-1}]$$

One can then see that any $\lambda_t \in [0,1]$ ensures that the resulting $\sigma^2_{\omega,t}$ is non-negative.

From the representation structure above, one can see that the forecast accuracy is described by λ_t. Given the posterior uncertainty from the previous period, $\Sigma_{t|t-1}$, the prior uncertainty is determined as follows:

$$\Sigma^m_{t|t} = \Sigma_{t|t-1} + (1-\lambda_t)\left(V[z_t] - \Sigma_{t|t-1}\right)$$

Uncertainty about z_t increases from $\Sigma_{t|t-1}$ to $\Sigma^m_{t|t}$ because prior knowledge is imperfectly represented in the new forecasts.

The value of λ_t is determined by the accuracy constraint (B.6). Given the optimal structure of $m_{i,t}$, the mutual information between $m_{i,t}$ and $m_{i,t-1}, n_{i,t-1}$ equals $\mathcal{I}\left(\tilde{m}_{i,t};\, z_{i,t|t-1}\right)$. Then, we can pin down λ_t as a function of ϕ_m:

$$\lambda_t = 1 - \phi_m$$

One can see that noisier memory corresponds to lower λ_t and higher prior uncertainty.

4 Cognitive Noise and Biased Forecasts

In this section, I show that forecasts based on the mental representation exhibit forecast biases found in Coibion and Gorodnichenko (2015) and Bordalo, Gennaioli, Ma, and Shleifer (2020b). I illustrate how we can interpret these biases through the proposed model. The model also provides an estimation strategy to infer the extent of cognitive

constraints from the survey forecasts.

4.1 Forecasts Subject to Cognitive Constraints

DM's time-t prior belief about z_t is derived as follows:

$$z^m_{i,i,t|t} = (1-\lambda)\, E[z_t] + \lambda\, z_{i,i,t|t-1} + \tilde{\omega}_{i,t}$$

We can see that forecasts are sluggish to incorporate past knowledge because memory is noisy ($\phi_m > 0$). When processing internal information is costly, remembered knowledge about z_t is anchored toward the default prior ($E[z_t]$). In the case of perfect memory, $z_t|\, m_{i,t}$ equals $z_t|\, m_{i,t-1}, n_{i,t-1}$.

Conditional on this prior belief, the posterior belief evolves according to the following formula:

$$z_{i,i,t|t} \;=\; (1-\kappa)\, z^m_{i,i,t|t} + \kappa\, z_t + \kappa\, \tilde{\nu}_t + \tilde{u}_{i,t}$$

We can see that forecasts are sluggish to track the current economy when subject to noisy news. When processing external information is costly, forecasts put less weight on new information and therefore are slow to catch up with new developments in z_t.

Combining these two formulas, beliefs about z_t follow the following law of motion:

$$z_{i,i,t|t} = (1-\lambda)\,(1-\kappa)\, E[z_t] + \lambda\,(1-\kappa)\, z_{i,i,t|t-1} + \kappa\, z_t + (1-\kappa)\,\omega_{i,t} + \kappa\, u_{i,t} \qquad \text{(D.11)}$$

The above equation summarizes the features of forecasts subject to cognitive noise. Because of noisy news, DM sluggishly recognizes a change in z_t. Because of noisy memory, DM sluggishly incorporates her past knowledge. And the idiosyncratic cognitive noise from noisy news and noisy memory creates forecast dispersion.

It is helpful to discuss how the noisy-memory assumption changes the predictions of

the traditional noisy-information model. If memory is perfect, then beliefs about z_t evolve according to the following formula:

$$z_{i,i,t|t} = (1 - \kappa^*)\, z_{i,i,t|t-1} + \kappa^*\, z_t + \kappa^*\, u_{i,t} \tag{D.12}$$

Comparing (D.11) to this law of motion, we can see three changes. With noisy memory, (1) prior knowledge receives a smaller weight ($\lambda < 1$), (2) new information receives a bigger weight ($\kappa \leq \kappa^*$), and (3) a new source of cognitive noise appears.

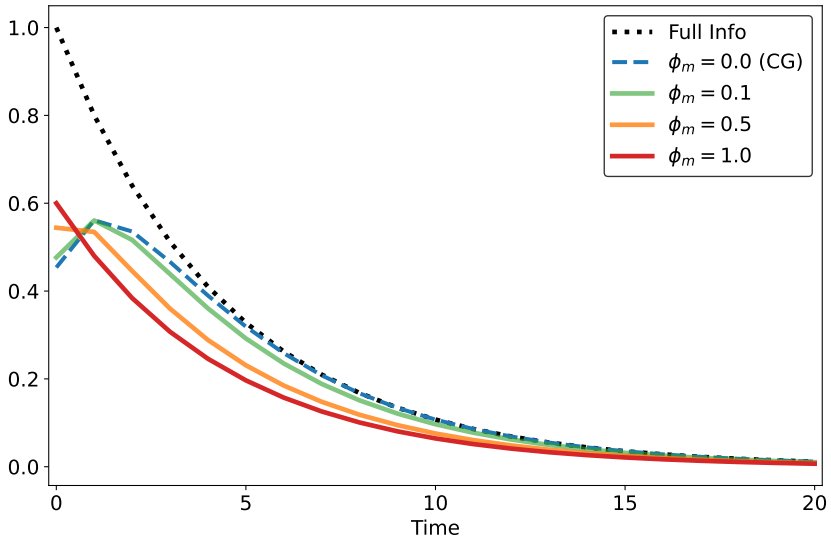
Impulse response function. Figure 3.1 illustrates the effects of noisy memory when learning about z_t. For this numerical exercise, I use the parameter values $\rho = 0.8$ and $\sigma_\epsilon^2 = 1.0$ for the data-generating process. I fix the extent of noisy news at $\phi_n = 0.4$.

The top panel shows the impulse response to innovation in z_t. The black dashed line shows the response of z_t. Other lines show the response of forecasts of z_t for varying degrees of noisy memory ϕ_m. The blue line is the perfect-memory case: As DM slowly learns about z_t, her forecasts undershoot the true z_t. With enough learning opportunities, the undershooting disappears, and forecasts closely follow the true y_t. In comparison, the red line is the no-memory case, in which DM has no access to her prior knowledge. Two features stand out. First, the initial response is more significant than the blue line. This is because the Kalman gain is higher when memory is imperfect. And second, learning is slow. Since DM cannot tap into her prior knowledge, learning takes a long time, even with the large Kalman gain. The other colored lines show the in-between cases, and the same intuition applies.

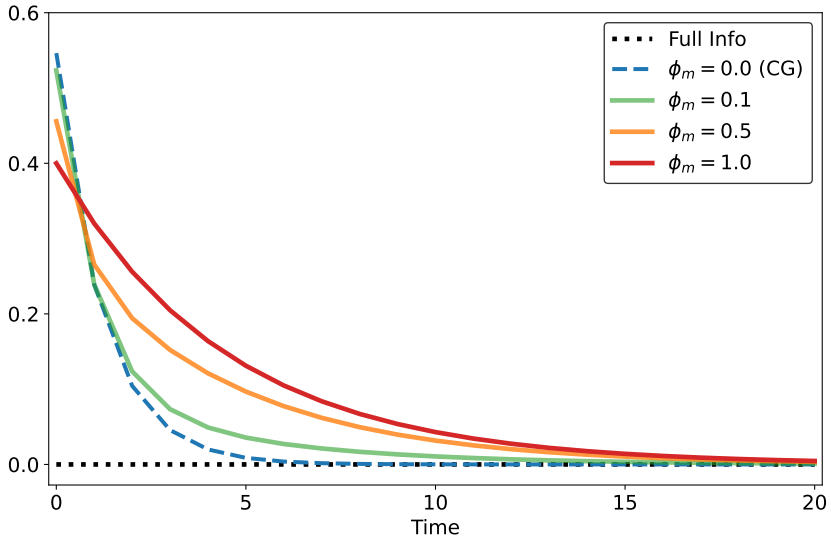
The bottom panel shows the impulse responses of the forecast errors, defined as $z_t - z_{i,t|t}$. When memory is perfect (the blue line), the initially large, positive response diminishes as learning accumulates. When memory is noisy, the forecast errors are initially smaller but remain large even as learning opportunities accrue.

Figure 3.1: Impulse response of forecasts

(a) Forecast



(b) Forecast errors



The figures show the impulse response to an innovation in z_t. The top panel shows the response of z_t and the forecast of z_t. The bottom panel shows the response of the forecast errors, defined as $z_t - z_{i,t|t}$. The data-generating process is described by $\rho = 0.8$ and $\sigma_\epsilon^2 = 1.0$. I fix the extent of noisy news as $\phi_n = 0.4$. The black dashed line shows the full-information case of perfect news and memory. Lines with different colors assume a varying degree of noisy memory.

4.2 Biases in Survey Forecasts

In this section, I revisit two regression specifications that test whether survey forecasts deviate from FIRE. Then, I discuss what the test results can inform us about the extent of underlying cognitive noise.

Three building-block assumptions of FIRE. Before investigating the features of survey forecasts, it is helpful to clarify the three assumptions embedded in FIRE. First, forecasters efficiently use all available information at hand. Thus, errors in forecasts are not systematically predictable by any element in the information set. Second, forecasters can access their prior knowledge perfectly. This means that forecast revisions should be in each individual forecaster's information set. Third, forecasters have access to the same complete information.

The first two assumptions predict that an econometrician cannot predict errors that an individual forecaster will make based on the latter's recent forecast revisions. The three assumptions together predict that an econometrician cannot predict errors in the average forecasts based on recent revisions in average forecasts.

Coibion and Gorodnichenko (2015) Regression Specification

Coibion and Gorodnichenko (2015) propose the following regression specification as a joint-hypothesis test for the three FIRE assumptions:

$$y_{t+h} - y_{i,t+h|t} = \alpha_C + \beta_C \left(y_{i,t+h|t} - y_{i,t+h|t-1} \right) + e_{t+h|t} \tag{D.13}$$

Here, $y_{i,t+h|t}$ and $y_{i,t+h|t-1}$ are the average forecasts of $y_{i,i,t+h|t}$ and $y_{i,i,t+h|t-1}$.

The authors find a positive β_C for many macroeconomic variables and reject the null hypothesis. They argue that relaxing the full-information assumption can explain the result. Intuitively, if the population does not have access to complete information, revisions

in the average forecasts will be sluggish, as at least some people make forecasts based on outdated information. Then, on average, forecasters revise their view about the future sluggishly in response to a change in the economy, and forecast errors are positively correlated with forecast revisions. Furthermore, the authors argue that a larger estimate of β_C can be interpreted as evidence for more significant information frictions.

The expectation-formation model introduced in Section 2 gives new insight into interpreting the regression coefficient.

Proposition 4. *For forecasts subject to cognitive noise, the asymptotic limit of* β_C *is*

$$\beta_C = \frac{1-\kappa}{\kappa}\left\{1 + (1-\lambda)\frac{\lambda(1-\kappa)\rho^2}{1-\lambda(1-\kappa)\rho^2}\right\}$$

if $\sigma_\nu^2 \to 0$. *Furthermore,* β_C *has the following properties:*

1. $\beta_C > 0$ *if* $\phi_n > 0$, *and* $\beta_C = 0$ *if* $\phi_n \to 0$.
2. $\frac{\partial \beta_C}{\partial \phi_n} > 0$, *and* $\frac{\partial \beta_C}{\partial \phi_m} < 0$ *if* $\phi_n \leq \bar{\phi}_n \equiv \bar{g}(\rho, \sigma_\epsilon^2)$.

Proof. See Appendix 10.4. □

Because of noisy news, the proposed model generates a positive β_C. Forecasters update their beliefs sluggishly because they do not have perfect awareness of the current state. As discussed in the previous section, the Kalman gain κ of less than one captures such sluggishness. In addition, noisier news generates a smaller gain and a larger β_C, as argued in Coibion and Gorodnichenko (2015).

A new insight from the proposed model is that noisy memory and noisy news jointly determine the Kalman gain. With noisier memory, the recalled prior knowledge is less accurate. Since uncertainty about the state is higher, forecasters put a larger weight on incoming data, which results in a higher Kalman gain and a lower β_C.

Bordalo, Gennaioli, Ma, and Shleifer (2020b) Regression Specification

Bordalo, Gennaioli, Ma, and Shleifer (2020b) propose the following regression specification as a joint-hypothesis test for the first two FIRE assumptions:

$$y_{t+h} - y_{i,i,t+h|t} = \alpha_I + \beta_I \left(y_{i,i,t+h|t} - y_{i,i,t+h|t-1} \right) + e_{i,t+h|t} \tag{D.14}$$

The authors reject the null hypothesis and find a negative β_I for many macroeconomic variables in contrast to the result from Coibion and Gorodnichenko (2015). They propose a non-Bayesian expectation model to explain the negative coefficient. The main idea is that forecasters irrationally put too much weight on new observations and over-revise their forecasts. Based on such a model, forecasters are not using available information efficiently, which generates a nonzero β_I. Furthermore, the authors argue that a more negative estimate of β_I can be interpreted as the extent of irrationality.

In contrast, I propose to relax the perfect-memory assumption while keeping the Bayesian-efficiency assumption. The proposed model offers an alternative interpretation of the regression coefficient as follows.

Proposition 5. *For forecasts subject to cognitive noise, the asymptotic limit of β_I is*

$$\beta_I = -\frac{(1-\lambda)(1-\kappa)}{2(1-\lambda)(1-\kappa) + \rho^{-2} - 1}$$

if $\rho > 0$. Furthermore, β_I has the following properties.

1. *$\beta_I < 0$ if $\phi_m > 0$, and $\beta_I = 0$ if $\phi_m \to 0$.*
2. *$\frac{\partial \beta_I}{\partial \phi_n} < 0$, and $\frac{\partial \beta_I}{\partial \phi_m} < 0$.*

Proof. See Appendix 10.4. □

The regression coefficient captures the bias in underusing past information. Because of

noisy memory, forecasts put less weight on past knowledge, which is captured by negative β_I.

Furthermore, noisy news and noisy memory jointly determine this forecast bias. Noisier memory leads to more underuse of past information, which generates a more negative β_I. With noisier news, forecasters rely more on their memory when making forecasts. Since external information is less effective in correcting the bias, β_I is more negative.

Identification of the Extent of Cognitive Constraints

From Propositions 1 and 2, we can see that the two regression coefficients can pin down the severity of noisy news and noisy memory.

Lemma 2. *Given levels of β_C and β_I identify a unique pair of ϕ_n and ϕ_m, if it exists.*

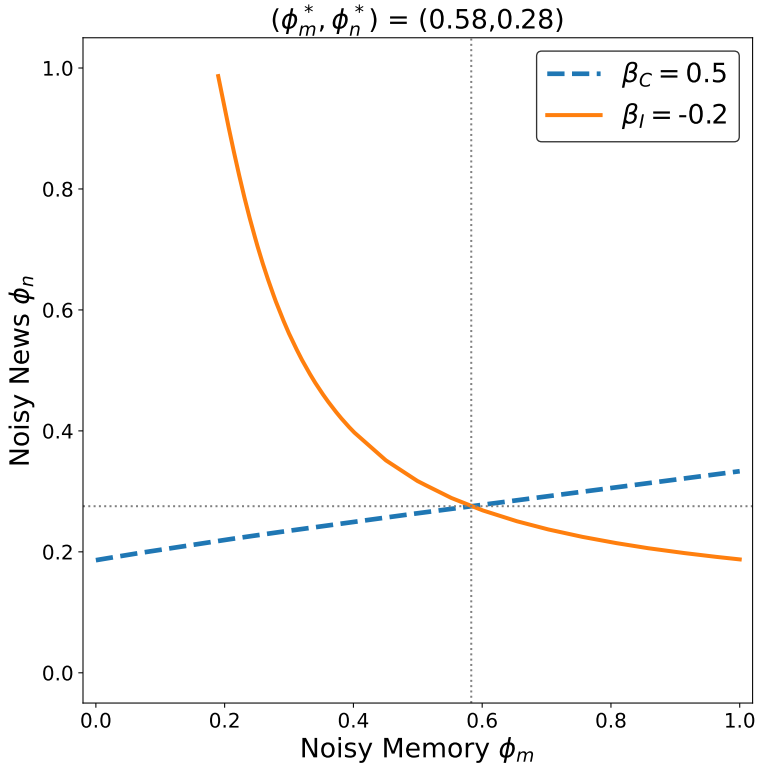
Proof. We can find the pairs of ϕ_n and ϕ_m that generate given levels of β_C and β_I (that is, the iso-curve). The iso-curve for β_C is upward-sloping, and the iso-curve for β_I is downward-sloping. Therefore, if the two iso-curves cross, they only cross once. □

Figure 3.2 illustrates the lemma. I assume that $\rho = 0.8$ and $\sigma_\epsilon^2 = 1.0$. The blue solid line is the iso-curve when $\beta_C = 0.5$. And the orange dashed line is the iso-curve when $\beta_I = -0.2$. We can see that the iso-curve for β_C is upward-sloping; more aggressive belief updating due to noisier memory is offset by more sluggish belief updating due to noisier news. We can also see that the iso-curve for β_I is downward-sloping; more underuse of past information due to noisier memory is offset if reliance on memory declines because of less noisy news. These two iso-curves cross once at most, identifying the extent of noisy news and noisy memory that can jointly predict the two estimated regression coefficients.

5 Extended Model

In Section 2, I assumed that DM is fully aware of the parameters generating z_t. In this section, I assume forecasters are also learning about the long-run mean of the forecast

Figure 3.2: β_C and β_I jointly identify the extent of cognitive noise



This figure shows the iso-curves for the two regression coefficients in (D.13) and (D.14). The blue solid line displays the pairs of noisy-news constraint ϕ_n and noisy-memory constraint ϕ_m that generate $\beta_C = 0.5$. The orange dashed line displays such pairs that generate $\beta_I = -0.2$. The point at which the two lines cross is the estimated extent of noisy news and noisy memory — that is, $\phi_n^* = 0.58$ and $\phi_m^* = 0.28$. The data-generating process is described by $\rho = 0.8$ and $\sigma_\epsilon^2 = 1$.

variable. I show that this extension can improve the model predictions in explaining the features of long-run forecasts.

5.1 Learning about the Long Run

Before I estimate the model, I revisit a commonly made assumption in the literature: that people are perfectly aware of the model. It is often motivated by the idea that people adapt to their environment and learn to make optimal economic decisions. However, I show this assumption is not innocuous in the proposed model. As discussed in Azeredo da

Silveira et al. (2020), when prior knowledge is imperfectly accessed, forecasters do not reach complete awareness of the model parameters.

One aspect of the environment that is particularly important for making long-horizon forecasts is the mean of the forecast process. Therefore, I assume that DM does not know the exact level of μ and has to learn about it, starting from a Gaussian prior:

$$\mu \sim \mathcal{N}(\bar{\mu}, \Omega)$$

The state variable relevant for predicting future realizations is expanded from y_t to (μ, y_t). This is because forecasts for y_{t+h} depend on DM's beliefs about μ and y_t. I denote this state vector as

$$x_t = \begin{pmatrix} \mu \\ z_t \end{pmatrix}.$$

All other assumptions are the same as in Section 2.

5.2 The Optimal Cognitive Process

The optimization problem for deriving the optimal cognitive process is the same as described in Section 2; the optimal process minimizes the objective function (B.1) subject to the information environment (B.3), (B.4), (B.5), and (B.6). However, the optimal cognitive process differs from the one introduced in Section 3 because the state variables are multivariate. In this section, I sketch the optimal cognitive process. Detailed derivations are in Appendix 10.3.

Implications of the linear-Gaussian structure. We can see that the initial prior about x_t is Gaussian. Therefore, the linear-Gaussian structure of noisy news and noisy memory again ensures that DM's belief about x_t follows a Gaussian distribution. DM's beliefs about

x_t (based on her cognitive state) are described with the following notation:

$$x_\tau |\, m_{i,t} \sim \mathcal{N}\left(x^m_{i,i,\tau|t},\, \Sigma^m_{\tau|t}\right)$$
$$x_\tau |\, m_{i,t}, n_{i,t} \sim \mathcal{N}\left(x_{i,i,\tau|t},\, \Sigma_{\tau|t}\right)$$

The loss function. The loss function reduces to

$$\sum_{t=0}^{\infty} \beta^t \, trace\left(\Sigma_{t|t}\, Q\right).$$

Q is a matrix defined as $Q \equiv \sum_{h=1}^{H} \alpha_h \, \alpha_h'$, where $\alpha_h = \begin{pmatrix} 1-\rho^h & \rho^h \end{pmatrix}$.

Optimal representation of noisy news. I first derive the optimal structure of the noisy news $n_{i,t}$. The optimal representation of N_t, denoted as $\tilde{n}_{i,t}$, takes the form

$$\tilde{n}_{i,t} = \tilde{K}_t \cdot E[x_t |\, N_t] + \tilde{u}_{i,t}$$

for some matrix $\tilde{K}_t$ and idiosyncratic noise $\tilde{u}_{i,t} \sim \mathcal{N}\left(O,\, \Sigma_{u,t}\right)$. The structure is similar to the optimal $n_{i,t}$ in Section 3. Since the forecast accuracy depends on the posterior uncertainty about x_t, the optimal summary of the information in N_t is captured by $E[x_t |\, N_t]$.

Under the assumed structure of the external news N_t in (B.2), the optimal $\tilde{K}_t$ and $\sigma_{u,t}$ are determined as follows:

$$\tilde{K}_t = \kappa_t \cdot \frac{\Sigma^m_{t|t}\, e\, e'}{e'\, \Sigma^m_{t|t}\, e}$$
$$\Sigma_{u,t} = \sigma^2_{u,t} \cdot \frac{\Sigma^m_{t|t}\, e\, e'\, \Sigma^m_{t|t}}{\left(e'\, \Sigma^m_{t|t}\, e\right)^{-2}}$$

κ_t and $\sigma^2_{u,t}$ were derived in Section 3. The vector $e' = \begin{pmatrix} 0 & 1 \end{pmatrix}$ picks out z_t from the state vector x_t. The above expression shows that the information represented in $\tilde{n}_{i,t}$ is $E[z_t |\, N_t]$

(with random errors). This is because the information in N_t about the additional state variable μ is subsumed in $E[z_t|\, N_t]$.

Optimal representation of noisy memory. The optimal representation of $(m_{i,t-1}, n_{i,t-1})$ is described with $\tilde{m}_{i,t}$ such that

$$\tilde{m}_{i,t} = \tilde{\Lambda}_t \cdot x_{i,i,t|t-1} + \tilde{\omega}_{i,t}$$

for some matrix $\tilde{\Lambda}_t$ and idiosyncratic noise $\tilde{\omega}_{i,t} \sim \mathcal{N}(O,\, \Sigma_{\omega,t})$. Intuitively, it is optimal to represent knowledge about x_t from the internal information $(m_{i,t-1}, n_{i,t-1})$, which is summarized as $E[x_t|\, m_{i,t-1}, n_{i,t-1}]$. I apply the normalization so that the accuracy of the representation is entirely determined by $\tilde{\Lambda}_t$:

$$Cov\left[x_{i,i,t|t-1}, \tilde{m}_{i,t}\right] = V[\tilde{m}_{i,t}]$$

This pins down the memory-noise variance $\Sigma_{\omega,t}$ as a function of $\tilde{\Lambda}_t$.

$$\Sigma_{\omega,t} = \left(I - \tilde{\Lambda}_t\right) V[x_{i,t|t-1}] \tilde{\Lambda}_t'$$

Any $\tilde{\Lambda}_t$ is feasible as long as the resulting $\Sigma_{\omega,t}$ is a proper variance-covariance matrix (that is, symmetric and positive-semidefinite).

In the appendix, I describe how $\tilde{\Lambda}_t$ can be derived. The complication arises because the information constraint (B.6) cannot completely determine the noisy memory anymore. To see why, note that the constraint reduces to

$$\mathcal{I}\left(m_{i,t};\, m_{i,t-1}, n_{i,t-1}\right) = -\frac{1}{2} \det\left(I - \tilde{\Lambda}_t\right) \leq -\frac{1}{2} \ln \phi_n.$$

That is, this constraint limits the determinant of $I - \tilde{\Lambda}_t$, leaving the elements of $\tilde{\Lambda}_t$ to be specified.

When solving for $\tilde{\Lambda}_t$, I consider a myopic case, in which $\beta \to 0$. I first define a matrix Γ_t that is crucial for determining the $\tilde{\Lambda}_t$. I call this matrix a *memory-priority* matrix and define it as follows:

$$\Gamma_t = (I - K_{t+1})' Q (I - K_{t+1})$$

The matrix Γ_t roughly captures how some information receives higher priority than other information. Two matrices show why I make such interpretations. First, the elements in $(I - K_{t+1})$ would be large if external information oes not resolve much uncertainty about the state, in which case a more accurate memory would be helpful. Second, the matrix Q is from the loss function of incorrect forecasting. If some elements in Q were high, more accurate memory would be helpful.

I show in the appendix that $V[x_{i,i,t+1|t}]^{\frac{1}{2}} \Gamma_t V[x_{i,i,t+1|t}]^{\frac{1}{2}}$ can be eigen-decomposed to $U_t G_t U_t'$, where U_t is an orthonormal matrix storing the eigenvectors and G_t is a diagonal matrix storing eigenvalues in descending order (that is, $g_{1,t} > g_{2,t}$). Then, the optimal Λ_t satisfies

$$\tilde{\Lambda}_t = V[x_{i,i,t+1|t}]^{\frac{1}{2}} U_t D_t U_t' V[x_{i,i,t+1|t}]^{-\frac{1}{2}},$$

where a diagonal matrix D_t is defined to be

$$D_t = \begin{cases} \begin{pmatrix} 1 - \left(\frac{g_{2,t}}{g_{1,t}}\phi_m\right)^{\frac{1}{2}} & 0 \\ 0 & 1 - \left(\frac{g_{1,t}}{g_{2,t}}\phi_m\right)^{\frac{1}{2}} \end{pmatrix} & \text{if} \quad \phi_m < \frac{g_{2,t}}{g_{1,t}} \\ \begin{pmatrix} 1 - \phi_m & 0 \\ 0 & 0 \end{pmatrix} & \text{otherwise.} \end{cases}$$

One can easily see that $\det(I - \Lambda_t) = \phi_m$. The derivation above shows how the rank of the memory variable is determined. The first case is when the dimension of the remembered knowledge is not reduced (that is, $m_{i,t+1}$ is two-dimensional). In this case, the first

diagonal element in D_t is higher than the second one, indicating that the corresponding orthogonalized factor receives a higher weight. The second case is when memory stores information in $x_{i,i,t+1|t}$ in a lower dimension. The first diagonal element in D_t receives the biggest possible weight satisfying the memory constraint, while the second element is zero.

Summary. We have seen the derivation for optimal noisy news and noisy memory, which is described by the sequence of $\{K_t, \Sigma_{u,t}, \Lambda_t, \Sigma_{\omega,t}\}_{t=0}^{\infty}$. The time-$t$ prior belief is described with $\tilde{\Lambda}_t$:

$$x^m_{i,t|t} = x_{i,t|t-1} + \left(I - \tilde{\Lambda}_t\right)\left(E[x_t] - x_{i,t|t-1}\right) + \tilde{\omega}_{i,t}$$
$$\Sigma^m_{t|t} = \Sigma_{t|t-1} + \left(I - \tilde{\Lambda}_t\right)\left(V[x_t] - \Sigma_{t|t-1}\right)$$

And the posterior belief is described with $\tilde{K}_t$:

$$x_{i,i,t|t} = \left(I - \tilde{K}_t\right) x^m_{i,i,t|t} + \tilde{K}_t\, x_t + \tilde{\nu}_t + \tilde{u}_{i,t}$$
$$\Sigma_{t|t} = \left(I - \tilde{K}_t\right)\Sigma^m_{t|t}$$

Here, $\tilde{\nu}_t \sim \mathcal{N}(O,\, \Sigma_\nu)$, whose variance is defined as $\Sigma_\nu = \kappa_t^2 \left(e'\,\Sigma^m_{t|t}\,e\right)^{-2} \Sigma^m_{t|t}\, e\, e'\, \Sigma^m_{t|t}$.

5.3 Perpetual Uncertainty about the Long Run

This section briefly discusses how DM learns about the long-run mean when she is subject to cognitive noise. Based on this discussion, I show the model predictions about the forecast-error-revision test for different forecast horizons.

When DM can access her internal information perfectly, she has complete access to all the past noisy news. In this case, the subjective uncertainty about the mean is

$$V[\mu|\, n_{i,t}, n_{i,t-1}, \cdots, n_{i,0}] = \left(\Omega^{-1} + t \times c\right)^{-1},$$

where Ω is the prior variance about μ, and c is a constant. We can see that the precision of knowledge linearly increases in time; the uncertainty eventually converges to zero after a long learning period.

Noisy memory qualitatively changes this prediction as investigated in Azeredo da Silveira et al. (2020). If DM imperfectly accesses internal information, $V[\mu|\, m_{i,t}, n_{i,t}]$ does not converge to zero even after a long learning period. The intuition is straightforward: cognitive noise prevents forecasters from reaching complete awareness even after an infinitely long learning period.

Why does it matter that DM is imperfectly aware of the long-run mean? It matters because DM will continuously update her beliefs about the mean as new data come, although she correctly understands that the mean is a constant parameter. When y_t is high, the DM partly attributes it to higher-than-expected μ and expects future y_t to be persistently high. This prediction is similar to extrapolative-expectation models in the finance literature. My model implies that a limited memory might be the reason such extrapolation occurs.

Impulse response function. Figure 3.3 illustrates the effect of learning about the long run. I use the same data-generating process as Figure 3.1 and set the cognitive parameters as $\phi_n = 0.4$, $\phi_m = 0.1$, and $\Omega = 1$.

The top panel shows the impulse response to innovation in z_t. The black dashed line is the response of z_t. The blue line is the response of forecasts for z_t. As in Figure 3.1, learning about y_t is sluggish because of noisy news.[3] The orange line shows the forecast for μ. As discussed earlier, DM perceives that z_t is high partly because the long-run mean is high and revises her belief about μ upward.

The bottom panel of Figure 3.3 displays the response of four-quarter-ahead forecasts for varying degrees of Ω. I realign the lines to compare forecasts to the realized z_{t+4}. We can see whether forecasts undershoot or overshoot compared to the black dashed line. We

[3]We can see that the impulse response of z_t more closely tracks z_t in Figure 3.3 than in Figure 3.1. This is because uncertainty about the long run increases uncertainty about z_t, pushing up the Kalman gain.

see initial undershooting for all values of Ω because of the noisy news. However, forecasts start overshooting after a few periods for some Ω. When Ω is high, DM revises her beliefs about the long-run mean too much, which offsets the undershooting due to noisy news. In this case, the forecast errors, defined as $z_{t+4} - z_{i,t+4|t}$, are initially positive in response to innovation in z_t but soon turn negative. This prediction is consistent with findings in Angeletos et al. (2021). The authors analyze the professional forecasters' year-ahead forecasts for unemployment and inflation and their impulse response to a specific shock series constructed by Angeletos et al. (2020).

Error-revision regression. The perpetual uncertainty about the long run also implies that the regression coefficients in the forecast error-revision test (D.13) and (D.14) will not be constant for different forecast horizons.

Consider the regression coefficient applied to forecasts for μ. Denoting the mean forecasts as $\hat{\mu}_{i,t} \equiv E[\mu|\, m_{i,t}, n_{i,t}]$ and the average forecasts as $\hat{\mu}_t \equiv \int \hat{\mu}_{i,t} d\, i$, we can see that

$$\beta_C^\mu = \frac{Cov\left[\mu - \hat{\mu}_t, \hat{\mu}_t - \hat{\mu}_{t-1}|\, \mu\right]}{V[\hat{\mu}_t - \hat{\mu}_{t-1}|\, \mu]} = -\frac{1}{2}$$
$$\beta_I^\mu = \frac{Cov\left[\mu - \hat{\mu}_{i,t}, \hat{\mu}_{i,t} - \hat{\mu}_{i,t-1}|\, \mu\right]}{V[\hat{\mu}_{i,t} - \hat{\mu}_{i,t-1}|\, \mu]} = -\frac{1}{2}.$$

The derivation is straightforward. We can deduce that $\beta_C = -\frac{V[\hat{\mu}_t|\mu] - Cov[\hat{\mu}_t, \hat{\mu}_{t-1}|\mu]}{2(V[\hat{\mu}_t|\mu] - Cov[\hat{\mu}_t, \hat{\mu}_{t-1}|\mu])}$ and must equal $-\frac{1}{2}$. The same reasoning applies to β_I.[4] Forecasters revise their views about μ although μ is a fixed parameter.

Figure 3.4 illustrates the model predictions for β_C and β_I for varying forecast horizons. I fix the degree of noisy news and noisy memory at levels in Figure 3.2 that generate the targeted β_C and β_I. I use $\Omega = 0.2$; this level corresponds to the posterior variance of μ if DM had access to twenty years of data. The figure shows that both coefficients become more negative for longer forecast horizons. As shown earlier, for forecasts far enough ahead, β_C^μ and β_I^μ are close to $-\frac{1}{2}$.

[4] Derivations for other horizons are in Appendix 10.6.

The pattern in Figure 3.4 is in line with empirical findings in the literature. d'Arienzo (2020) and Wang (2021) analyze professional forecasters' projections of interest rates. Both authors find that longer-horizon forecasts feature more negative biases when the regressions (D.13) and (D.14) are estimated. Bordalo, Gennaioli, Porta, and Shleifer (2019) and Bordalo, Gennaioli, La Porta, and Shleifer (2020a) find a similar pattern for stock analysts' forecasts for companies' long-term earnings.

6 Estimating the Extent of the Cognitive Constraints

In this section, I estimate the two cognitive constraints by using professional forecasters' survey data.

6.1 Data

Survey forecast data are from the Survey of Professional Forecasters (SPF), administered by the Federal Reserve Bank of Philadelphia. Once every quarter, around forty forecasters (mostly from academia and banks) participate in this survey. The earliest survey started in 1968. I use survey forecasts made until the second quarter of 2022.

Among the survey questions, those in the section titled "The U.S. Business Indicators" ask forecasters to submit their views about aspects of the overall US economy, which include output, price level, labor and housing markets, and cost of borrowing. I investigate whether the proposed model can explain features of survey forecasts made for that section.[5] Table 3.1 lists the variables.

For data on the time series of macroeconomic variables, I use the Real-Time Data Set from the Federal Reserve Bank of Philadelphia whenever possible. This data set provides the history of data releases for each variable. Since the variables in the National Income and Product Accounts are often redefined or reclassified, the final data release (that is, the most recently available data) often does not include the same variables forecast by the

[5]There have been some categorical changes, as the survey forms changed over time, but I include eleven variables that are consistently included most of the time.

professional forecasters in the data set. Therefore, I compare the initial releases of each variable to the corresponding SPF forecasts.

6.2 Estimation Strategy

I estimate four parameters that affect how DM makes forecasts about the macroeconomic variables: ϕ_n and ϕ_m (the severity of the two cognitive limitations), σ_ν^2 (the amount of correlated noise), and Ω (the unconditional prior uncertainty about the long-run mean).[6] The parameters describing the data-generating process are estimated from the realized macroeconomic variables. I assume that each variable is described as a univariate autoregressive process. Related parameters are in Appendix 10.5. Finally, I assume that the longest forecast horizon of the loss function (B.1) is eight quarters ahead since the SPF asks forecasters to submit their forecasts for up to two years ahead for the "The U.S. Business Indicators" section.

I transform the survey forecast data so that the unit of forecasts is the log difference from the previous quarter for most variables. I use change from the previous quarter for the unemployment rate and the three financial variables in Table 3.1. Surveyed forecasters make projections for different horizons, so all forecasts are annualized to make the units consistent. I use forecasts up to four quarters ahead.

I drop some observations to restrict the influence of a few outlier variables. In each period, I remove forecasts if they are five quantiles outside the median level. I remove forecasters if they participate for fewer than ten periods. I further restrict samples to measure the forecast behavior in the normal business cycle. During periods of big swings in the macroeconomy such as the COVID-19 pandemic, it is likely that forecasters use different forecasting methods and therefore exhibit different behaviors. Since my model does not capture such structural changes, I use a simple algorithm to remove likely structural-

[6]Failure to consider the correlated noise ν_t can bias the model estimation. This is because the estimated regression coefficient β_C from (D.13) is attenuated when forecast noise is correlated among forecasters (Coibion and Gorodnichenko (2015) and Gemmi and Valchev (2021)).

change episodes. Namely, I compute the average size of forecast revisions among forecasters each period and remove the top five percentile periods. This procedure systematically identifies significant revision episodes, removing the beginning of the pandemic for unemployment but not for less affected variables.

Estimation targets. The first two data moments I use are the regression coefficients described earlier: β_C from (D.13) and β_I from (D.14). The forecast error-revision pair is available for the forecast horizon for up to three quarters. I estimate the regression by pooling the four forecast horizons.[7] For the individual-level regression, I include individual and horizon dummies to purge variations due to the fixed effects.

I panel-bootstrap the SPF individual-forecast data and build bootstrap samples of the targeted moments. Each sample contains on average forty individual forecasters, as in the survey data. The first two panels of Table 3.1 report this coefficient. The table reports the median and confidence interval of 5%–95% estimates. As discussed in Section 4.2, we see positive β_C and negative β_I across the variables. These two moments can identify the underlying degree of information constraints, given the two remaining parameters σ_ν^2 and Ω. I also report the OLS estimates in Table 3.3. The bootstrapped estimates and the OLS estimates are similar.

Two more moments are used to estimate the model. These moments are informative about σ_ν^2 and Ω. Based on Gemmi and Valchev (2021), I measure the size of Kalman gains using the following specification:

$$\left(y_{i,i,t+h|t} - y_{i,i,t+h|t-1}\right) - \left(y_{t+h|t} - y_{t+h|t-1}\right) = \alpha_K + \beta_K \left(y_{t+h|t-1} - y_{i,i,t+h|t-1}\right) + error_{i,t+h|t-1} \tag{F.15}$$

[7] I pool the different forecast horizons for two reasons. An obvious reason is to increase power. But more importantly, I am interested in estimating the constraints in processing information about the near-term economy, not just the current economy. A literal interpretation of the model is that DM gets news only about the current economy. (Since the time unit is a quarter, DM gets news about the current quarter only.) However, it would be realistic to assume that forecasters learn about the near-term economy.

This specification estimates how forecasters revise their views about the current economy in response to news about it. The strategy is to partial out the effects of the correlated noise by de-meaning individual forecasts. Since the correlated noise attenuates β_C, comparing the above regression coefficient to the Kalman gain implied by β_C is informative about the degree of correlated noise.[8]

I pool the forecast horizons and control for individual-forecaster and forecast-horizon fixed effects. The right panel of Table 3.1 reports the regression coefficient. The table reports the median and confidence interval of 5%–95% estimates, and the OLS estimates are in Table 3.3.

I use a similar specification to measure how long-term forecasts are revised in response to news about the near-term economy. We need frequent long-term forecast data to estimate this regression. The SPF collects these data for the Consumer Price Index (CPI) but not for other macroeconomic variables. I use forecasts for the annual average rate of headline CPI inflation over the next ten years to estimate the regression. The coefficient is estimated to be 0.0862, statistically significant at the 1% level, with a standard error of 0.0175. More details are in Appendix 10.5. Since data are not available to conduct a similar analysis for other macroeconomic variables, I target the estimated coefficient for all variables. While it is not feasible to verify the validity of this assumption, we can at least see that the estimated regression coefficient for (F.15) is broadly similar across variables.

6.3 Estimation Results

I now estimate parameters that fit each bootstrapped sample discussed in the previous section. I report the median estimate and the 5%–95% confidence band in Table 3.4. Figure 3.6 reports the estimates of noisy news ϕ_n and noisy memory ϕ_m.

In Section 4, I showed that the methodology in Coibion and Gorodnichenko (2015) underestimates the magnitude of ϕ_n because it misattributes the extra sensitivity from noisy

[8]The authors show that the new estimate of the Kalman gain is smaller for most macroeconomic variables they study.

memory to low ϕ_n. To investigate the extent of underestimation, I repeat the estimation procedure while assuming $\phi_m = 0$. In this case, I estimate two parameters, ϕ_n and σ_ν^2, that match the two estimation targets, β_C and β_K.

The top panel in Figure 3.6 compares ϕ_n estimated using the proposed model to that estimated assuming perfect memory. As expected, the estimated ϕ_n is larger with noisy memory. On average, the baseline ϕ_n is twice as large as ϕ_n estimated the using Coibion and Gorodnichenko (2015) methodology. The bottom panel illustrates the estimated ϕ_m. For most variables, ϕ_m is significant and positive. Overall, the estimated parameters are somewhat stable: the average levels are $\phi_n = 0.31$ and $\phi_m = 0.24$; the median levels are $\phi_n = 0.34$ and $\phi_m = 0.22$.

Table 3.2 assesses the model fit using the point estimate. The top and bottom panels show the targeted and untargeted moments, respectively. This table reports the average levels across macroeconomic variables.

We confirm that the model matches the targeted moments well. For untargeted moments, I show variations in forecasts and forecast revisions. For each variable, I report variations in the time series (that is, dispersion of the consensus forecasts) and in the cross section (that is, dispersion of the individual forecasts at any given time). All measures are the standard deviation scaled by the standard deviation of the forecast variable. We can see that the estimated model has a reasonable quantitative fit.

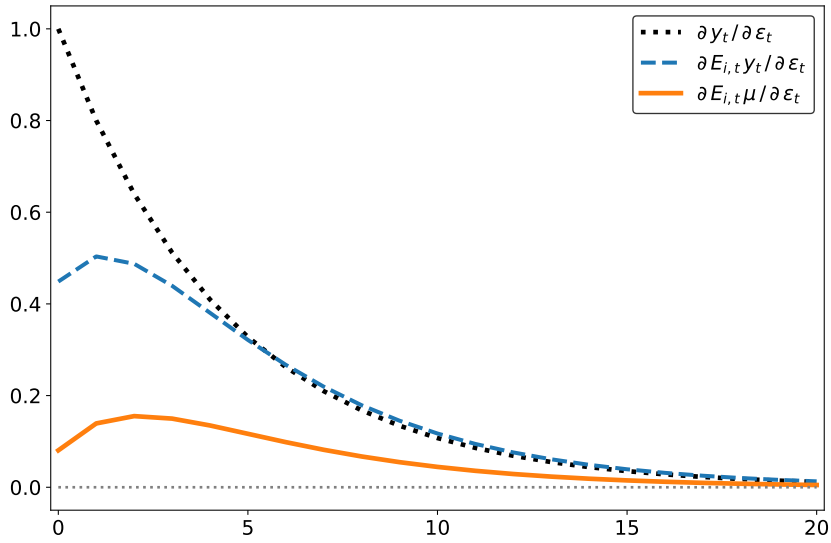
Figure 3.5 illustrates the fit of untargeted moments for all macroeconomic variables. Although the model is too stylized to replicate variations across macroeconomic variables perfectly, it generates a good fit. The detailed data for this figure are available in Table 3.6.

Table 3.1: Estimated regression coefficients

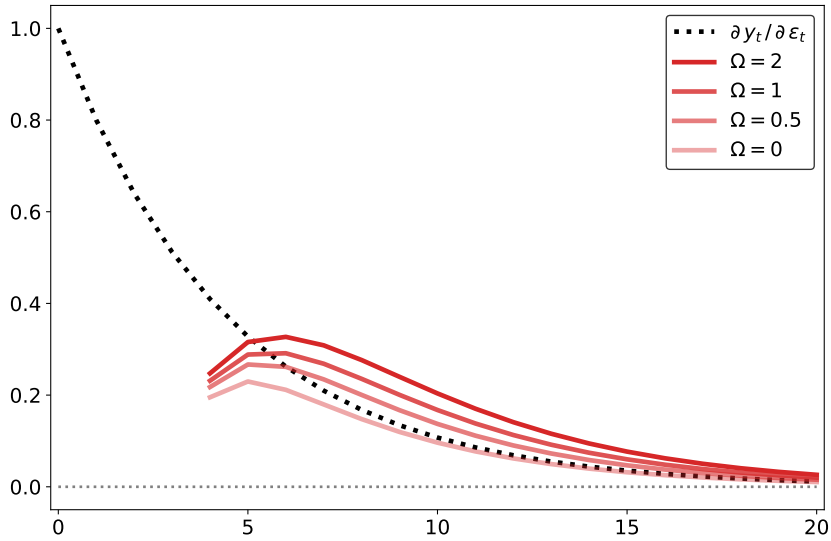
	β_C	CI	β_I	CI	β_K	CI
Nominal Gross Domestic Product	0.55	(0.43,0.66)	-0.27	(-0.3,-0.24)	0.55	(0.52,0.58)
Real Gross Domestic Product	0.36	(0.26,0.45)	-0.24	(-0.27,-0.21)	0.61	(0.58,0.63)
GDP Chain-Weighted Price Index	0.56	(0.43,0.69)	-0.32	(-0.36,-0.28)	0.6	(0.57,0.64)
Corporate Profits after Taxes	0.49	(0.32,0.66)	-0.44	(-0.48,-0.4)	0.51	(0.49,0.53)
Civilian Unemployment Rate	0.56	(0.51,0.62)	-0.05	(-0.08,-0.02)	0.63	(0.6,0.65)
Industrial Production Index	0.53	(0.44,0.61)	-0.18	(-0.22,-0.15)	0.57	(0.55,0.61)
Housing Starts	0.41	(0.31,0.49)	-0.27	(-0.32,-0.22)	0.58	(0.55,0.6)
Consumer Price Index	0.46	(0.32,0.61)	-0.17	(-0.22,-0.12)	0.57	(0.53,0.6)
Treasury Bill Rate, 3-month	0.28	(0.2,0.34)	-0.01	(-0.03,0.01)	0.73	(0.69,0.77)
AAA Corporate Bond Yield	0.03	(-0.03,0.09)	-0.35	(-0.38,-0.32)	0.68	(0.66,0.7)
Treasury Bond Rate, 10-year	0.26	(0.21,0.33)	-0.12	(-0.15,-0.1)	0.7	(0.67,0.73)

Figure 3.3: Impulse-response functions when learning about the long run

(a) Forecasts of z_t and μ

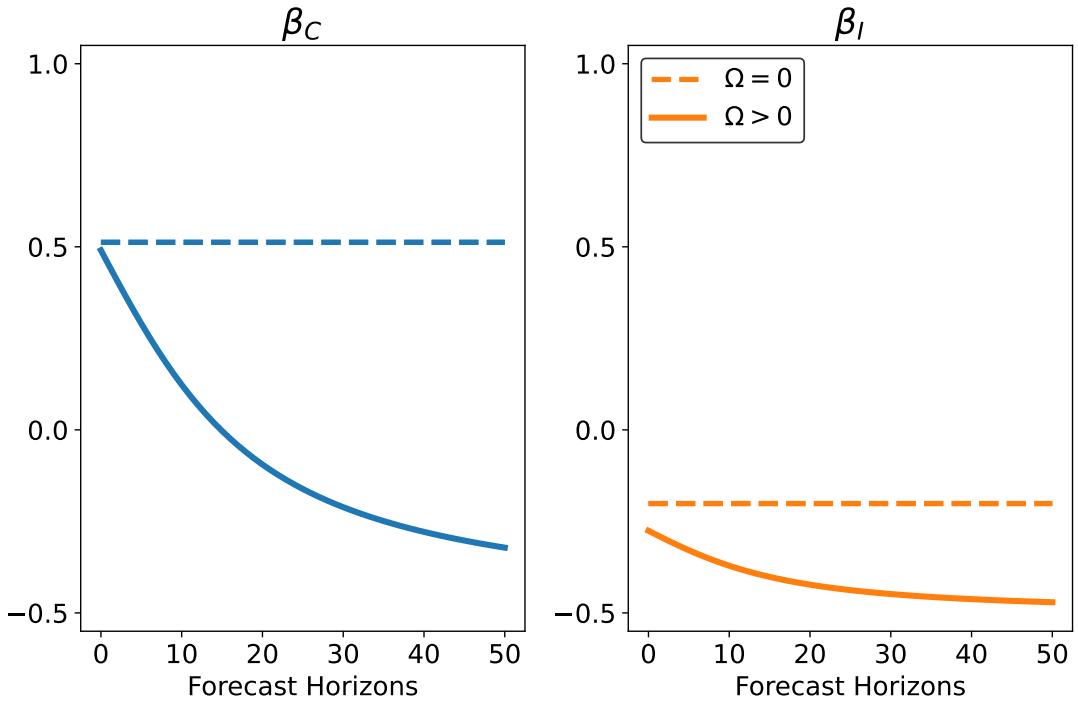


(b) Forecasts of z_{t+4}



The figures show the impulse response to an innovation in y_t. The data-generating process is described by $\rho = 0.8$ and $\sigma_\epsilon^2 = 1$. The top panel shows the response of y_t and the forecast of y_t and μ. I fix the cognitive noise as $\phi_n = 0.4$ and $\phi_m = 0.1$, and I set $\Omega = 1$. The bottom panel shows the response of four-period-ahead forecasts ($y_{i,t+4|t}$). Different lines assume varying degrees of Ω, the initial uncertainty about μ.

Figure 3.4: β_C and β_I when learning about the long run



This figure shows model predictions of the two regression coefficients in (D.13) and (D.14) for different forecast horizons. The extent of cognitive noise is from Figure 3.2: $\phi_n^* = 0.28$ and $\phi_m^* = 0.58$. The gray solid line is the model prediction when DM does not have to learn about the long run ($\Omega = 0$). The black dashed line is when DM learns about the long run ($\Omega = 1$). The data-generating process is described by $\rho = 0.8$ and $\sigma_\epsilon^2 = 1$.

Figure 3.5: Estimated β_C and β_I

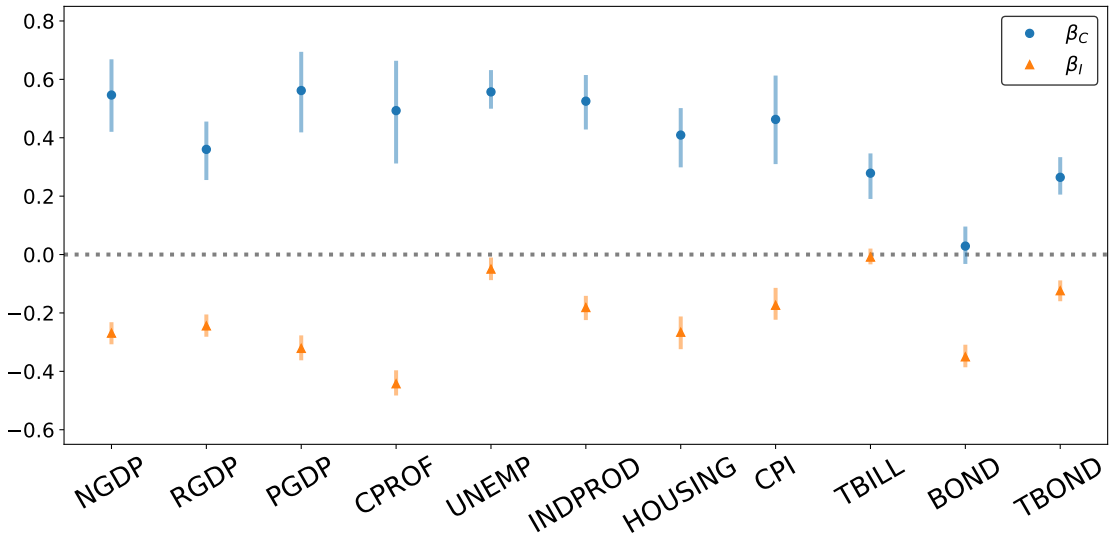
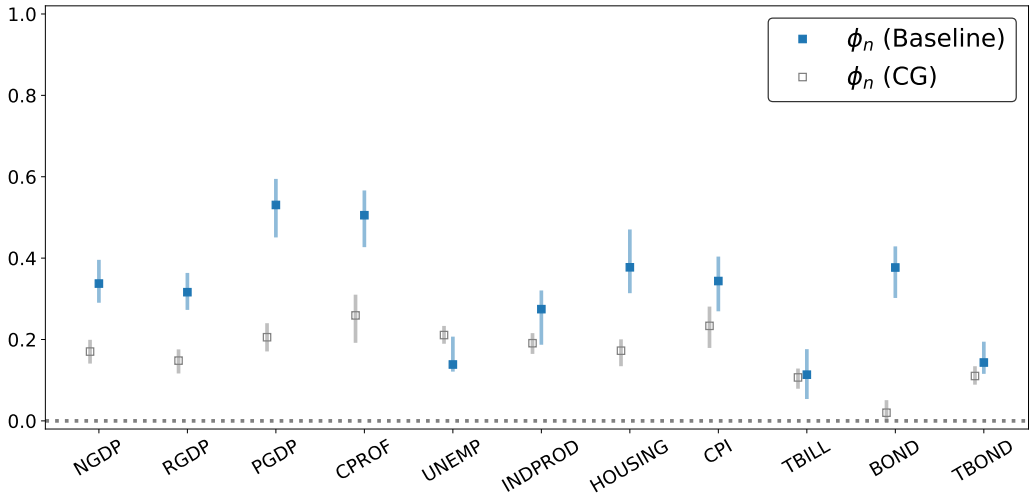


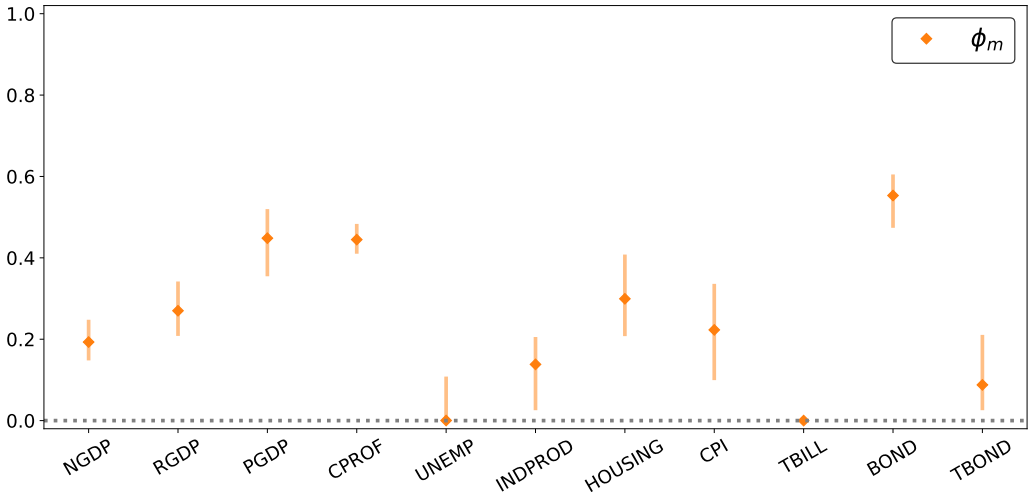
Table 3.1 reports the estimated regression coefficients. I study the variables in the SPF's "U.S. Business Indicators" section. From left to right, each panel presents the coefficients in (D.13), (D.14), and (F.15). The last two regressions include individual and horizon fixed effects. I panel-bootstrap the SPF data. The dot is the median estimate, and the error band shows the 5% and 95% estimates. The OLS estimates are reported in Table 3.3. Figure 3.5 visualizes the estimation of β_C and β_I, whose x-axis uses the abbreviated variable names that are in the same order as in the table above.

Figure 3.6: Estimated Parameters

(a) Noisy News



(b) Noisy Memory



This figure's full variable names are in Table 3.1. The top panel reports the estimated extent of noisy news (ϕ_n), and the bottom panel reports that of noisy memory (ϕ_m). Estimation targets the panel-bootstrapped moments discussed in Table 3.1. The dot is the median estimate, and the error band contains the 5% and 95% estimates. Table 3.4 reports the detailed numerical results. In the top panel, I compare the estimated ϕ_n (labeled as "Baseline") to the estimation achieved under the Coibion and Gorodnichenko (2015) assumption (labeled as "CG"). For the latter, I impose $\phi_m = 0$ and estimate two parameters (ϕ_n and σ_ν^2) that match two targets (β_C and β_K).

Table 3.2: Model Fit

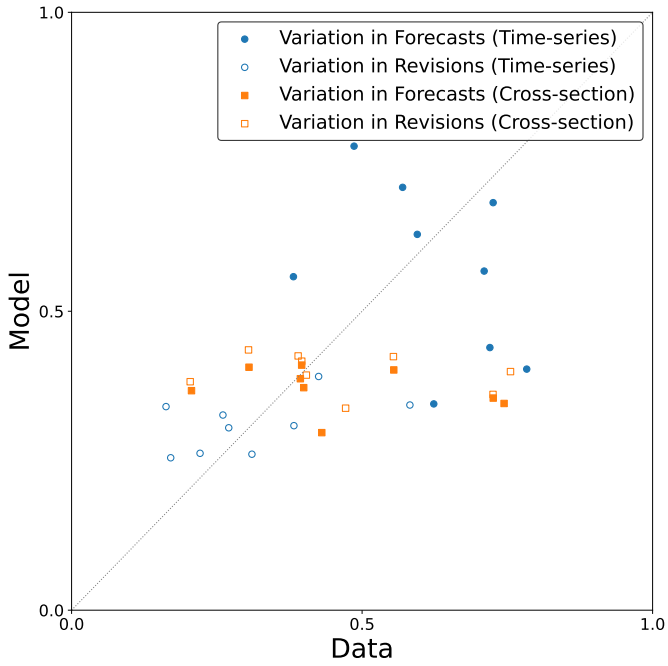
(a) Targeted moments (average across macroeconomic variables)

β_C		β_I		β_K		$\beta_{\mu,K}$	
Data	Model	Data	Model	Data	Model	Data	Model
0.41	0.41	-0.22	-0.2	0.61	0.61	0.08	0.07

(b) Not-targeted moments (average across macroeconomic variables)

Variation in Forecasts				Variation in Revisions			
Time Series		Cross Section		Time Series		Cross Section	
Data	Model	Data	Model	Data	Model	Data	Model
0.6	0.6	0.43	0.36	0.31	0.32	0.44	0.38

Figure 3.7: Not-targeted moments (all macroeconomic variables)



The tables evaluate the fit of the estimated model when using the median estimates in Table 3.4. The upper panel shows the targeted moments, and the lower panel shows untargeted moments. Both panels report the average value across all macroeconomic variables in Table 3.1. For untargeted moments, I report variations of forecasts and forecast revisions in the time series and cross sections, averaged across four consecutive forecast horizons (current to three quarters ahead). The unit of all measures is the standard deviation scaled by the standard deviation of the forecast variables. The figure illustrates the untargeted moments for all macroeconomic variables. The detailed data for this figure are available in Table 3.6.

7 An Illustrative Macroeconomic Model

In this section, I study the macroeconomic implications of the proposed expectation-formation model. Using a standard New Keynesian model, I show that if expectations are formed as in my model, inflation may be more variable, worsening the central bank's policy trade-off in stabilizing inflation and output. I discuss the efficient monetary policy in this environment and what harm can be done if the central bank conducts monetary policy that is only efficient under conventional expectation assumptions.

7.1 Firms' Decision Problem

Optimal Price Setting

Suppose firm i reconsiders its price $P_{i,t}$ in period t. The new price that it chooses maximizes the expected value of the firm's (current market value) profits. This pricing decision does not constrain any future decisions. Thus, it suffices to consider the effects of the choice on expected profits in those future states in which the price has not yet again been re-optimized. The firm's new price solves the following problem:

$$\max_{P_{i,t}} \quad E_{i,t}\left[\sum_{h=0}^{\infty} \alpha^h Q_{t,t+h}\left(P_{i,t}\, Y_{i,t+h|t} - \Psi_{t+h}\left(Y_{i,t+h|t}\right)\right)\right]$$

Here, α is the probability of not resetting prices, $Q_{t,t+h}$ is the stochastic discount factor for evaluating the future nominal payoffs generated at $t+h$, $Y_{i,t+h|t}$ is the output demanded in period $t+h$ if the price remains at the one chosen at time t, and Ψ_{t+h} is the (nominal) cost function at time $t+h$. Firm i takes into account that the demand $Y_{i,t+h|t}$ is given as

$$Y_{i,t+h|t} = \left(\frac{P_{i,t}}{P_{t+h}}\right)^{\eta} C_{t+h},$$

where η is the elasticity of substitution among goods, P_{t+k} is the aggregate price at time $t+h$, and C_{t+h} is the aggregate consumption at time $t+h$.

I use the notation $E_{i,t}$ to denote firm i's subjective expectation at time t. While in the conventional New Keynesian model $E_{i,t}$ refers to full-information rational expectations, I propose that the firm's expectations are formed according to the cognitive limitations proposed in earlier sections. The firm's objective depends only on aggregate conditions at the various dates $t+h$. Thus, under rational expectations, the optimal price $P^*_{i,t}$ would be the same for all i that reconsider their price at date t. However, under the expectation-formation model proposed in this paper, the optimal choice $P^*_{i,t}$ may differ across firms because of their differing expectations.

The firm's optimal price $P^*_{i,t}$ is derived using the first-order condition. Below I describe the first-order Taylor expansion of this condition around the zero-inflation steady state (I use lowercase to denote the log of the variable denoted in uppercase):

$$p^*_{i,t} - p_{i,t-1} = E_{i,t}\left[\sum_{h=0}^{\infty} (\alpha\beta)^h \left\{(1-\alpha\beta)\left(mc_{t+h} - mc\right) + \pi_{t+h}\right\}\right]$$

Here, mc_{t+h} is the log of real marginal cost at $t+h$ (mc is its steady-state value), and π_{t+h} is inflation at $t+h$ defined as $\log P_{t+h} - \log P_{t+h-1}$. As detailed in Appendix 10.7, the marginal costs do not depend on the quantity that a firm supplies. This is because of the assumed feature of the production function that the marginal product of labor does not depend on the quantity of production. Thus, firm i treats the nominal marginal costs as evolving independently of its own pricing decision; they only depend on aggregate variables that the firm takes as given.[9] Let us define

$$z_{t+h} \equiv (1-\alpha\beta)\left(mc_{t+h} - mc\right) + \pi_{t+h}. \tag{G.16}$$

Thus, the firm's expectations of the current and future z_t determine its subjectively optimal

[9] I introduce this assumption for the sake of simplicity. However, even when the firm's marginal product of labor varies with the quantity supplied, the subjectively optimal price will still depend only on its expectations about aggregate economic variables. See Gali (2008, Chapter 3).

price:

$$p^*_{i,t} - p_{i,t-1} = E_{i,t}\left[\sum_{h=0}^{\infty} (\alpha\beta)^h z_{t+h}\right] \tag{G.17}$$

7.2 Aggregate Economy

Real Marginal Costs

The real marginal costs are derived from the rest of the economy. As detailed in Appendix 10.7, the household optimization problem and market-clearing conditions imply that

$$mc_t - mc = \chi\, x_t + e_t. \tag{G.18}$$

χ depends on the elasticities of the consumption and labor utility functions, and x_t is defined as $y_t - y^e_t$, where y^e_t is the efficient level of output. Finally, e_t is the cost-push shock. While I do not take a stance on the source of cost-push shocks, one example is a time-varying, exogenous wage markup. The cost-push shock is a transitory i.i.d. shock fluctuating around zero.

Monetary Policy

Because of cost-push shocks, it is infeasible for the central bank to stabilize both inflation and the output gap fully. Thus, the central bank faces a policy trade-off in stabilizing the two variables. I assume that monetary policy is specified by a targeting rule of the form

$$x_t = -s\,\pi_t, \tag{G.19}$$

where s is a constant scalar that I later calibrate to match the relative variability of the output gap to inflation in the data.

The targeting rule illustrates the relationship between x_t and π_t that the central bank seeks to maintain in response to a fluctuation in the economy. The rule implies that the central bank accepts inflation higher than its long-run target (assumed to be zero in the model) if and only if there is a negative output gap. Likewise, the targeting rule requires inflation to be lower than the long-run target when there is a positive output gap at the same time. The implication of such a targeting rule for the path of interest rates can be derived using the household intertemporal optimization condition.

Aggregation

Once firms reconsider their price and choose their subjectively optimal price $P_{i,t}^*$, the aggregate price index is formed according to

$$P_t = \left[\alpha \left(P_{t-1}\right)^{1-\eta} + (1-\alpha)\left(P_t^*\right)^{1-\eta}\right]^{\frac{1}{1-\eta}},$$

where $P_t^* \equiv \int P_{i,t}^* \, di$ is the average reset price of firms that reconsider their prices at time t. The first-order Taylor expansion of the price index implies $\pi_t = (1-\alpha)\left(p_t^* - p_{t-1}\right)$. Therefore, we can derive the aggregate inflation by averaging the expectations of the different firms:

$$\pi_t = (1-\alpha)\bar{E}_t \left[\sum_{h=0}^{\infty} (\alpha\beta)^h \, z_{t+h}\right] \tag{G.20}$$

Here, $\bar{E}_t$ averages the expectations $E_{i,t}$ of all individual firms.

Determination of z_t

By substituting (G.18) and (G.19) into (G.16), we can deduce that z_t is determined as follows:

$$z_t = \left\{1 - (1-\alpha\beta)\,\sigma\, s\right\}\pi_t + (1-\alpha\beta)\, e_t \tag{G.21}$$

Equations (G.20) and (G.21) together imply that z_t is determined by firms' expectations about current and future z_t and the exogenous shock e_t. Thus, once we specify how firms forecast z_t, we have a complete theory of how inflation, the output gap, and z_t evolve.

7.3 Firms' Macroeconomic Expectations

Suppose that firms form their forecasts under the assumption that z_t is an i.i.d. process such that

$$z_t \sim \mathcal{N}\left(\mu,\, \sigma_z^2\right). \tag{G.22}$$

As discussed below, this assumption is correct under FIRE. As in the proposed expectation model, firms are not perfectly aware either of the current value of z_t or of the mean μ of the distribution from which it is drawn. (σ_z^2 is assumed to be known to DM.) Firms' prior beliefs about μ are described as

$$\mu \sim \mathcal{N}\left(0,\, \Omega\right)$$

for some positive Ω.

I denote the average beliefs of firms about z_t and μ as $\hat{z}_t$ and $\hat{\mu}_t$, respectively. Then, $\hat{z}_t$ and $\hat{\mu}_t$ have the following law of motion:

$$\hat{z}_t = \lambda\left(1-\kappa\right)\hat{\mu}_{t-1} + \kappa\, z_t \tag{G.23}$$

The average expectation about the mean is

$$\hat{\mu}_t = \lambda\left(1-\kappa_\mu\right)\hat{\mu}_{t-1} + \kappa_\mu\, z_t. \tag{G.24}$$

Firms' beliefs are influenced by the realized z_t, which are determined by the rest of the aggregate economy, including the monetary policy discussed in the following section.

7.4 Expectation Formations and Inflation Dynamics

From (G.20), we see that inflation is determined by the average expectations of firms about the current and future courses of z_t. As seen from (G.23) and (G.24), they are are completely specified by two state variables: $\hat{\mu}_{t-1}$ (the average belief about μ in the previous period) and the realized value of z_t. Furthermore, π_t and the exogenous shock e_t determine the evolution of z_t, as described in (G.21). Combining all these equations, we can deduce that the inflation process is a linear function of e_t and $\hat{\mu}_{t-1}$:

$$\pi_t = \varphi_e \, e_t + \varphi_\mu \, \hat{\mu}_{t-1} \tag{G.25}$$

We can see that π_t is a persistent process since $\hat{\mu}_{t-1}$ is a function of $z_{t-1}, z_{t-2}, \cdots, z_0$, which are in turn functions of lags of π_t and e_t. The coefficients φ_e and φ_μ are derived as

$$\varphi_e = \frac{\delta}{1 + \delta \, \sigma \, s + \frac{1}{\alpha} \frac{1-\hat{\kappa}}{\hat{\kappa}}}$$
$$\varphi_\mu = \frac{1}{1 + \delta \, \sigma \, s + \frac{1}{\alpha} \frac{1-\hat{\kappa}}{\hat{\kappa}}} \, \frac{1-\alpha}{\alpha} \, \frac{\hat{b}}{\hat{\kappa}},$$

where $\delta \equiv \frac{(1-\alpha)(1-\alpha\beta)}{\alpha}$, $\hat{\kappa} = \kappa + \kappa_\mu$, and $\hat{b} = \lambda(1-\kappa) + \frac{\alpha\beta}{1-\alpha\beta}\lambda(1-\kappa_\mu)$. See Appendix 10.7 for detailed derivation.

Comparison of different expectation assumptions. Different assumptions about expectation formation result in different inflation dynamics, as captured by $\hat{\kappa}$ and $\hat{b}$. I compare three cases: FIRE ($\phi_n = 0$ and $\phi_m = 0$), the conventional models of information frictions ($\phi_n > 0$ and $\phi_m = 0$), and finally the proposed expectation model ($\phi_n > 0$ and $\phi_m > 0$).

Under FIRE, firms are perfectly aware of z_t and μ. Therefore, firms expect the future marginal costs to be zero on average (since $\mu = 0$) and set their prices to match the current marginal costs. Therefore, the aggregate inflation is proportional to the realized z_t. The

inflation process is derived as follows:

$$\pi_t = \frac{\delta}{1 + \delta\,\sigma\,s}\,e_t$$

Under conventional models of information frictions, firms are imperfectly aware of z_t but have come to learn the true mean of the distribution z_t is drawn from. Thus, their subjectively optimal price is equal to the perceived value of the current marginal costs. This is because they correctly expect that their future marginal costs are zero on average. The inflation process is derived as

$$\pi_t = \frac{\delta}{1 + \delta\,\sigma\,s + \frac{1}{\alpha}\,\frac{1-\kappa^*}{\kappa^*}}e_t,$$

where κ^* refers to the Kalman gain when updating firms' belief about z_t under the perfect-memory assumption. Firms' reset prices are less responsive to the realized cost-push shocks than under FIRE. This is because firms are not perfectly aware of them when resetting prices. Accordingly, while aggregate inflation is still proportional to the cost-push shocks, the dependence is more muted.

Under the proposed model, the inflation process is derived as follows:

$$\pi_t = \rho_\mu\,\pi_{t-1} + \gamma_0\,e_t + \gamma_1\,e_{t-1}$$

Here, the coefficients on the cost-push shocks are derived as $\gamma_0 = \varphi_e + \varphi_m\kappa_\mu$ and $\gamma_1 = -\varphi_e\,\lambda\,(1-\kappa_\mu)$. Inflation is persistent, unlike in the previous two expectation models. This is because of the fluctuating beliefs about the long run, as the coefficient ρ_μ is the serial correlation of $\hat{\mu}_t$.

7.5 Calibration

I now discuss how I choose the model parameters. The parameters describing the expectation process come from the previous estimation section. I take the median estimates across macroeconomic variables. For the baseline model, I use $\phi_n = 0.34$, $\phi_m = 0.22$, and $\Omega/\sigma_y^2 = 0.12$. For the conventional models of information frictions, I use $\phi_n = 0.15$.

I set $\chi = 2$ to reflect that the elasticity coefficients of the consumption and labor utility function are both one, following the discussion in Hazell et al. (2022). I assume that firms discount their future revenues with $\beta = 0.99$ because I consider the time unit of the model to be a quarter. The frequency of price changes is matched to the slope of the Phillips curve estimated in the literature. The inflation response to a 1% increase in the output gap (holding the expectation terms) is estimated to be 0.024 in Rotemberg and Woodford (1997) and 0.0062 in Hazell et al. (2022). I target 0.01 as a midpoint.

Finally, I pin down s in the central bank's targeting rule (G.19) and the variance of the cost-push shock σ_e^2 to match the empirical volatility of inflation and the output gap. I use the quarterly log changes of the CPI for inflation. For the output gap, I use the difference between the log of real gross domestic product (RGDP) and the log of potential RGDP. All data are from Federal Reserve Economic Data (FRED). The standard deviation of the CPI is 0.35% per quarter, and the standard deviation of the output gap is 2.48% per quarter.

7.6 Monetary Policy and Inflation Variability

We have seen that the expectation-formation process shapes inflation dynamics. In this section, I consider the effects of alternative monetary policies on inflation variability and the role of expectation formation. To do so, I consider values of s in (G.19) given by

$$s = s^* \frac{\theta}{1-\theta}, \tag{G.26}$$

where s^* is the calibrated value of s. Thus, $\theta = \frac{1}{2}$ represents the typical monetary policy, bringing the model-predicted volatility of inflation and output closer to the data.

The strength of inflation targeting is measured by $\theta \in [0, 1]$. Complete inflation stabilization is captured by $\theta = 1$. In this case, in response to inflationary pressures from the cost-push shock, the central bank drives output far below the efficient level to stabilize inflation.

The top left panel in Figure 3.8 shows firms' subjective uncertainty about the long run. The x-axis corresponds to the strength of inflation targeting. I discuss the prediction for three different expectation assumptions: FIRE (black dotted line), the conventional models of information frictions (blue solid line), and the baseline model (orange solid line). As discussed earlier, firms are perfectly aware of the long run under FIRE and the conventional information-frictions model for any monetary-policy rule, but this prediction changes when noisy memory is also present. Firms continually feel uncertain about the long run and keep revising their views. In particular, the strength of inflation targeting matters; more stable inflation means more stable marginal costs, so firms become less uncertain about the long-run mean.

The top right panel of Figure 3.8 displays the inflation variability for a given monetary-policy rule on the x-axis. I confirm that stronger inflation targeting stabilizes the inflation process for all expectation assumptions. Furthermore, we can see that conventional information-friction models predict more stable inflation than under FIRE. Since firms are not perfectly aware of the realized marginal cost, they do not reflect it in their prices. In the baseline model, firms are imperfectly aware of both the realized marginal cost and its long-run mean. Therefore, their expectations of future marginal costs fluctuate, inducing more price fluctuations.

The bottom panel in Figure 3.8 illustrates the central bank's trade-off in simultaneously stabilizing inflation and the output gap. Under the conventional information-frictions model, the policy frontier shifts inward compared to FIRE; the economy faces less variable

inflation at any output variability. In the baseline model, the policy frontier shifts out, indicating that for any output variability, the economy bears more variable inflation.

7.7 Efficient Inflation Targeting

We have seen that the effect of monetary policy on inflation variability varies with the expectation assumptions. In this section, I study the efficient level of inflation targeting that maximizes social welfare for each expectation assumption.

Let us assume that social welfare depends on how variable the output gap and inflation are. Let us further assume that the welfare-relevant measure of the output gap is the output gap scaled by $\frac{1}{s^*}$. Thus, the welfare losses from the output gap and inflation are roughly comparable in size. Thus:

$$\mathcal{L} = (1-\omega)\ V[\tilde{x}_t] + \omega\, V[\pi_t] \tag{G.27}$$

Here, $\tilde{x}_t = \frac{1}{s^*}\, x_t$, and ω reflects the central bank's preference for stabilizing inflation over stabilizing the output gap. I find the optimal level of θ that minimizes the loss function.[10]

The left panel in Figure 3.9 displays the efficient weight for a given ω (the welfare weight on inflation). Under the conventional information-frictions assumption, it is efficient to put less emphasis on inflation targeting than under FIRE. Since the inflation process is less responsive to fluctuations in marginal cost, the central bank can put more weight on stabilizing the output gap. In comparison, putting more weight on inflation is efficient in the baseline model. Since the volatile inflation process feeds into more widely fluctuating beliefs about the long-run economy, the central bank prioritizes stabilizing inflation.

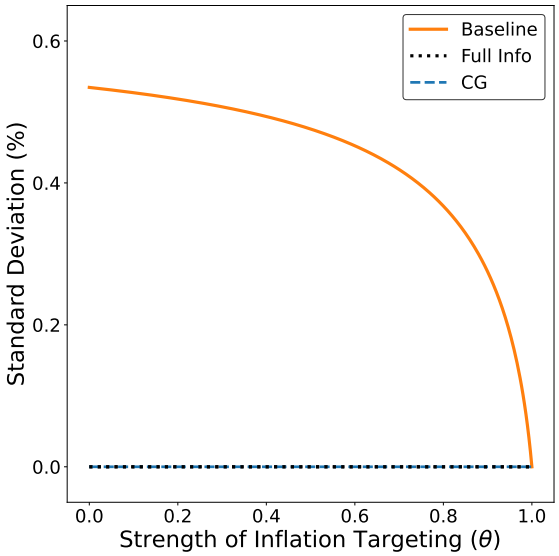
The right panel in Figure 3.9 illustrates that conducting monetary policy based on a

[10] One can consider the welfare-loss function whose measure of the output gap is not scaled. I propose to use the scaled output gap to see the effect of different expectation assumptions more clearly for the entire range of ω. In the current exercise, the standard deviation of the output gap is more than seven times larger than that of inflation. Thus, the efficient strength of inflation targeting is quite small unless the welfare weight on inflation is sizable.

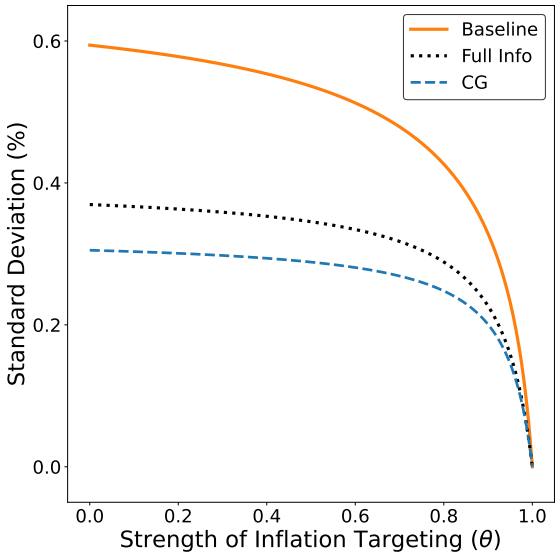
correct expectation assumption is essential. I show the additional inflation variability that the economy incurs if the central bank adopts a monetary policy that is only efficient under different expectation assumptions. The increased volatility is especially sizable when the central bank intends to produce more stable inflation (that is, when the welfare weight on inflation is high). That is, the central bank can generate volatile inflation because it is not cognizant that fluctuation in marginal costs will unanchor long-run expectations.

Figure 3.8: The effect of monetary policy

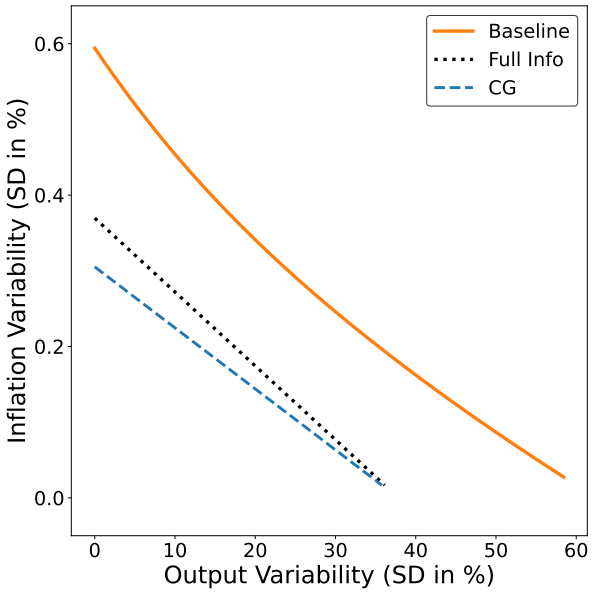
(a) Uncertainty about long run



(b) Inflation variability



(c) Stabilization trade-off

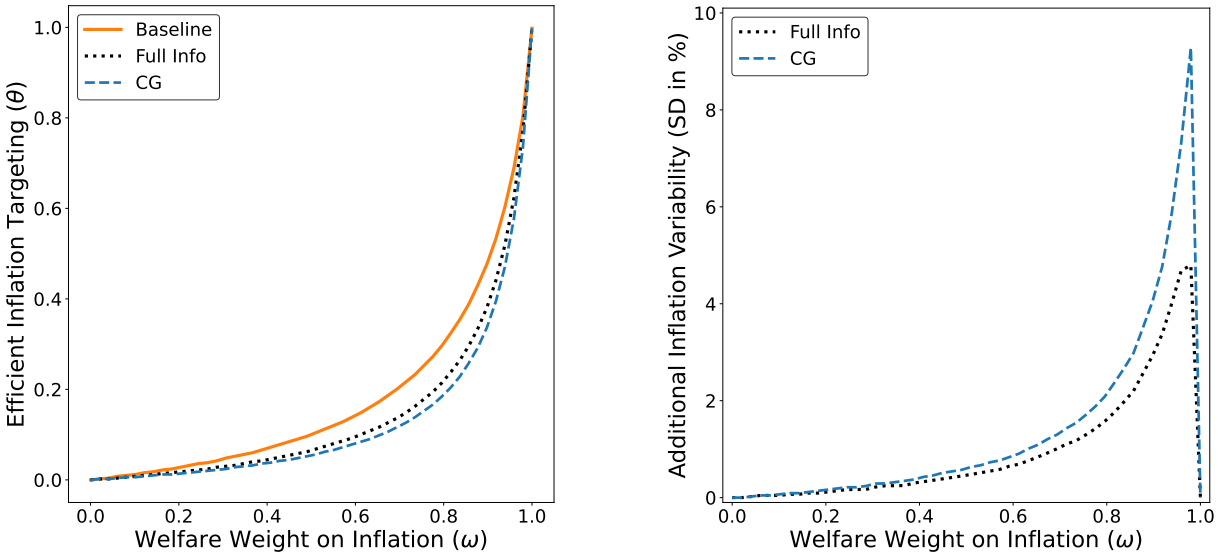


The figures above illustrate the macroeconomic dynamics for varying degrees of strength of inflation targeting (θ). For all figures, three lines correspond to different expectation-formation assumptions: "Baseline" is the proposed model, "Full Info" is the full-information model, and "CG" is the conventional models of information frictions. For each targeting rule θ on the x-axis, the top left panel displays the uncertainty about the long-run mean μ, and the top right panel shows the inflation variability. The bottom panel reports the policy trade-off between inflation stabilization and output-gap stabilization. The model parameters are stated in the main text.

Figure 3.9: Efficient policy

(a) Efficient inflation targeting

(b) Increased inflation volatility from naive policy



The left panel shows the efficient level of inflation targeting that minimizes the welfare loss (G.27). The welfare weight on inflation variability (ω) is on the x-axis. Three lines correspond to different expectation-formation assumptions as in Figure 3.8. The right panel shows the increased inflation variability from implementing inefficient targeting rules. The targeting rules are only efficient if expectations are not subject to noisy memory.

8 Conclusion

I proposed an expectations model in which economic agents make forecasts subject to information frictions. The proposed model accounts for puzzling patterns that conventional information-friction models cannot. It also offers an estimation strategy to identify the extent of information frictions. Using professional forecasters' overall projections of the US economy, I showed that an influential methodology previously proposed in the literature underestimates the extent of information frictions by half. Using the estimated model, I discussed the model's implications for inflation expectations and monetary policy. The public's expectations about the long-run state of the economy are not as well anchored as conventional information-friction models predict. I showed that the central bank's emphasis on inflation stabilization can be more desirable.

To reach these findings, I proposed that the relevant information friction is the cognitive constraint in processing the vast amount of information people have access to. Importantly, I proposed that economic agents process information both external and internal to their minds. This contrasts with conventional information-friction models, which implicitly assume that internal information is perfectly accessible. I showed that jointly considering the two information constraints is crucial to correctly estimating the extent of information frictions. To study the macroeconomic implications, I introduced the proposed expectation model into a standard New Keynesian model. I showed that price-setting firms have unanchored expectations about the long run when internal information is not perfectly accessible. Furthermore, I showed that policies that are efficient under conventional information-friction models generate excessive inflation volatility.

An important lesson from my analysis is that it is crucial to identify the fundamental bottleneck that keeps economic agents from making forecasts consistent with FIRE. I showed that finite capacity to process information — both external and internal — explains various features of survey forecasts that previous expectation-formation models can-

not. Recognition of these constraints allows one to see that conventional assumptions in macroeconomic models may not be well grounded. One example is the assumption that agents would be well aware of the long-run economic trends if the economy were stable. In the proposed model, agents' long-run expectations perpetually fluctuate even after extensive learning opportunities. This has the crucial implication that seemingly anchored long-run inflation expectations can start moving when agents witness bouts of high inflation. Thus, a monetary authority whose policies rely on the prospect of firmly anchored expectations can lose its grip on the economy, leaving economic agents to doubt the authority's ability to manage inflation. Empirically relevant expectation-formation models can guide the complex considerations that conducting monetary policy requires, especially in new environments yet to be experienced and analyzed.

9 Accompanying Tables and Figures

Table 3.3: Estimated regression coefficients using OLS

	β_C	SE	p-value	β_I	SE	p-value	β_K	SE	p-value
Nominal Gross Domestic Product	0.63	0.11	0.0	-0.27	0.04	0.0	0.54	0.03	0.0
Real Gross Domestic Product	0.45	0.12	0.0	-0.25	0.05	0.0	0.6	0.02	0.0
GDP Chain-Weighted Price Index	0.71	0.11	0.0	-0.32	0.04	0.0	0.6	0.03	0.0
Corporate Profits after Taxes	0.68	0.17	0.0	-0.44	0.05	0.0	0.51	0.03	0.0
Civilian Unemployment Rate	0.62	0.08	0.0	-0.05	0.05	0.31	0.62	0.02	0.0
Industrial Production Index	0.61	0.13	0.0	-0.18	0.06	0.0	0.58	0.02	0.0
Housing Starts	0.5	0.12	0.0	-0.25	0.06	0.0	0.58	0.02	0.0
Consumer Price Index	0.55	0.15	0.0	-0.17	0.09	0.04	0.56	0.03	0.0
Treasury Bill Rate, 3-month	0.29	0.05	0.0	-0.01	0.04	0.85	0.73	0.03	0.0
AAA Corporate Bond Yield	0.05	0.07	0.48	-0.35	0.04	0.0	0.68	0.02	0.0
Treasury Bond Rate, 10-year	0.28	0.07	0.0	-0.12	0.05	0.01	0.7	0.03	0.0

The first column shows the variables included in the SPF's "U.S. Business Indicators" section. The first panel displays the estimated regression coefficient from (D.13). The standard errors are robust to the presence of arbitrary heteroskedasticity and autocorrelation. The second panel shows the regression coefficient estimates from (D.14) when individual forecasts are pooled. For this regression, the standard errors are two-way clustered by forecasters and survey date. The last panel reports the regression coefficient from (F.15). The standard errors are similarly clustered two-way. I include individual and horizon fixed effects for the last two regression specifications.

Table 3.4: Estimated parameters

	ϕ_n	CI	ϕ_m	CI	Ω/σ_y^2	CI	σ_ν^2/σ_y^2	CI
Nominal Gross Domestic Product	0.34	(0.29,0.39)	0.19	(0.15,0.24)	0.16	(0.14,0.18)	0.14	(0.09,0.22)
Real Gross Domestic Product	0.32	(0.28,0.36)	0.27	(0.21,0.34)	0.12	(0.11,0.14)	0.23	(0.15,0.35)
GDP Chain-Weighted Price Index	0.53	(0.46,0.59)	0.45	(0.36,0.52)	0.1	(0.09,0.11)	0.52	(0.25,1.0)
Corporate Profits after Taxes	0.51	(0.43,0.56)	0.44	(0.41,0.48)	0.08	(0.08,0.09)	1.0	(1.0,1.0)
Civilian Unemployment Rate	0.14	(0.13,0.2)	0.0	(0.0,0.1)	1.0	(0.14,1.0)	0.0	(0.0,0.03)
Industrial Production Index	0.27	(0.19,0.32)	0.14	(0.03,0.2)	0.19	(0.14,1.0)	0.1	(0.07,0.16)
Housing Starts	0.38	(0.32,0.47)	0.3	(0.21,0.4)	0.12	(0.1,0.14)	0.32	(0.18,0.64)
Consumer Price Index	0.34	(0.27,0.4)	0.22	(0.1,0.33)	0.12	(0.09,0.27)	0.28	(0.1,0.58)
Treasury Bill Rate, 3-month	0.11	(0.06,0.17)	0.0	(0.0,0.0)	0.27	(0.16,1.0)	0.09	(0.03,0.39)
AAA Corporate Bond Yield	0.38	(0.31,0.42)	0.55	(0.48,0.6)	0.08	(0.08,0.09)	1.0	(0.61,1.0)
Treasury Bond Rate, 10-year	0.14	(0.12,0.19)	0.09	(0.03,0.21)	0.26	(0.12,0.92)	0.08	(0.03,0.13)

Each panel shows the estimated parameter and its confidence interval. The standard error is computed by panel-bootstrapping the SPF individual-forecast data. I report the 5% and the 95% point estimates from the bootstrapped samples. I estimate the scaled value of σ_ν^2 and Ω (divided by the variance of the forecast variable). I restrict these scaled σ_ν^2 and Ω to be between zero and one.

Table 3.5: Estimated parameters using Coibion and Gorodnichenko (2015) approach

	ϕ_n	CI	σ_ν^2/σ_y^2	CI
Nominal Gross Domestic Product	0.17	(0.15,0.19)	0.05	(0.02,0.08)
Real Gross Domestic Product	0.15	(0.12,0.17)	0.07	(0.04,0.12)
GDP Chain-Weighted Price Index	0.21	(0.18,0.24)	0.02	(0.0,0.06)
Corporate Profits after Taxes	0.26	(0.2,0.31)	0.17	(0.08,0.27)
Civilian Unemployment Rate	0.21	(0.19,0.23)	0.01	(0.0,0.03)
Industrial Production Index	0.19	(0.17,0.21)	0.04	(0.02,0.08)
Housing Starts	0.17	(0.14,0.2)	0.09	(0.05,0.13)
Consumer Price Index	0.23	(0.18,0.28)	0.11	(0.04,0.19)
Treasury Bill Rate, 3-month	0.11	(0.08,0.12)	0.03	(0.0,0.06)
AAA Corporate Bond Yield	0.02	(0.0,0.05)	0.16	(0.13,0.2)
Treasury Bond Rate, 10-year	0.11	(0.09,0.13)	0.05	(0.02,0.08)

Each panel shows the estimated parameter and its confidence interval. The standard error is computed by panel-bootstrapping the SPF individual-forecast data. I report the 5% and the 95% point estimates from the bootstrapped samples. I estimate the scaled value of σ_ν^2 (divided by the variance of the forecast variable). I restrict these scaled σ_ν^2 to be between zero and one.

Table 3.6: Model fit

(a) Targeted Moments

	β_C		β_I		β_K		$\beta_{\mu,K}$	
	Data	Model	Data	Model	Data	Model	Data	Model
Nominal Gross Domestic Product	0.55	0.54	-0.27	-0.27	0.55	0.55	0.08	0.08
Real Gross Domestic Product	0.36	0.35	-0.24	-0.24	0.61	0.61	0.08	0.08
GDP Chain-Weighted Price Index	0.56	0.57	-0.32	-0.32	0.6	0.6	0.08	0.08
Corporate Profits after Taxes	0.49	0.5	-0.44	-0.27	0.51	0.57	0.08	0.08
Civilian Unemployment Rate	0.56	0.59	-0.05	-0.05	0.63	0.65	0.08	0.04
Industrial Production Index	0.53	0.52	-0.18	-0.18	0.57	0.58	0.08	0.08
Housing Starts	0.41	0.4	-0.27	-0.27	0.58	0.58	0.08	0.08
Consumer Price Index	0.46	0.47	-0.17	-0.17	0.57	0.57	0.08	0.08
Treasury Bill Rate, 3-month	0.28	0.28	-0.01	-0.01	0.73	0.63	0.08	0.03
AAA Corporate Bond Yield	0.03	0.02	-0.35	-0.34	0.68	0.68	0.08	0.08
Treasury Bond Rate, 10-year	0.26	0.25	-0.12	-0.11	0.7	0.7	0.08	0.08

(b) Not-targeted moments

	Variation in Forecasts				Variation in Revisions			
	Time Series		Cross Section		Time Series		Cross Section	
	Data	Model	Data	Model	Data	Model	Data	Model
Nominal Gross Domestic Product	0.73	0.68	0.4	0.41	0.22	0.26	0.39	0.43
Real Gross Domestic Product	0.59	0.63	0.39	0.39	0.26	0.33	0.4	0.42
GDP Chain-Weighted Price Index	0.78	0.4	0.31	0.41	0.17	0.25	0.3	0.44
Corporate Profits after Taxes	0.62	0.35	0.74	0.35	0.31	0.26	0.73	0.36
Civilian Unemployment Rate	0.51	0.75	0.3	0.33	0.23	0.34	0.31	0.37
Industrial Production Index	0.57	0.71	0.4	0.37	0.27	0.31	0.4	0.39
Housing Starts	0.71	0.57	0.55	0.4	0.38	0.31	0.55	0.42
Consumer Price Index	0.38	0.56	0.21	0.37	0.16	0.34	0.2	0.38
Treasury Bill Rate, 3-month	0.54	0.78	0.29	0.25	0.39	0.39	0.34	0.27
AAA Corporate Bond Yield	0.72	0.44	0.73	0.35	0.58	0.34	0.76	0.4
Treasury Bond Rate, 10-year	0.49	0.78	0.43	0.3	0.43	0.39	0.47	0.34

The table compares the predictions of the estimated model to the data moments. The upper panel shows the targeted moments, and the lower panel shows untargeted moments. For untargeted moments, I report variations of forecasts and forecast revisions in the time series and cross sections, whose units are standard deviations scaled by the standard deviation of the forecast variables. For each macroeconomic variable, these moments are averaged across four consecutive forecast horizons (current to three quarters ahead).

10 Appendix

10.1 Derivation of the Optimal Cognitive Process

For any given state vector x_t, I show the optimal structure of the cognitive process, described by the sequence of $\{K_t, \sigma_{u,t}, \Lambda_t, \sigma_{\omega,t}\}_{t=0}^{\infty}$, that minimizes the loss function (B.7) subject to the information environment (B.3), (B.4), (B.5), and (B.6).

Proof: The Optimal Structure for the Representation

I show below that the optimal $n_{i,t}$ and $m_{i,t}$ are one-dimensional. In particular, I show that the optimal $n_{i,t}$ records $E[z_t|\, N_t]$ with noise while the optimal $m_{i,t}$ stores $z_{i,i,t|t-1}$ with noise.

Step 1: Partition of $n_{i,t}$ and $m_{i,t}$

Partition of $n_{i,t}$ We can partition $n_{i,t} = K_t \cdot N_t + u_{i,t}$ into the following form

$$\begin{pmatrix} \vec{n}_{i,t} \\ \\ \tilde{n}_{i,t} \end{pmatrix} = \begin{pmatrix} K_{a,t} & K_{b,t} \\ \\ K_{c,t} & K_{d,t} \end{pmatrix} \begin{pmatrix} \vec{N}_t \\ \\ E[x|\, N_t] \end{pmatrix} + \begin{pmatrix} \vec{u}_{i,t+1} \\ \\ \tilde{u}_{i,t+1} \end{pmatrix} \tag{J.28}$$

Note that the elements of $\vec{N}_t$ are not correlated with $E[x_t|\, N_t]$ and that $\vec{N}_t$ and $E[x_t|\, N_t]$ span the same vector space as N_t. I also impose the following normalization assumption

$$E[x_t|\, m_{i,t}, n_{i,t}] = \tilde{n}_{i,t} + cons \cdot E[x_t|\, m_{i,t}]$$

This relationship holds if and only if $E[x_t|\, N_t] - \tilde{n}_{i,t}$ is uncorrelated with all the elements in $n_{i,t}$ conditional on $m_{i,t}$. That is, the two requirements are

$$Cov\left[x_t - \tilde{n}_{i,t}, \vec{n}_{i,t}|\, m_{i,t}\right] = \vec{O} \tag{J.29a}$$

$$Cov\left[x_t - \tilde{n}_{i,t}, \tilde{n}_{i,t}|\, m_{i,t}\right] = O \tag{J.29b}$$

We can see that (J.29b) implies

$$\begin{aligned} &Cov\left[x_t, K_{d,t}\, E[x_t|\, N_t]|\, m_{i,t}\right] = V[K_{c,t}\, \vec{N}_t + K_{d,t}\, E[x_t|\, N_t] + \tilde{u}_{i,t}\Big|\, m_{i,t}] \\ \Leftrightarrow \quad &V[K_{c,t}\, \vec{N}_t + \tilde{u}_{i,t}\Big|\, m_{i,t}] = Cov\left[x_t, K_{d,t}\, E[x_t|\, N_t]|\, m_{i,t}\right] - K_{d,t}\, V[E[x_t|\, N_t]|\, m_{i,t}]\, K_{d,t}' \end{aligned}$$

The feasible set of $K_{d,t}$ is defined as $K_{d,t}$ that yields the right-hand-side term to be a proper variance-covariance matrix (that is, symmetric and p.s.d.).

Partition of $m_{i,t}$ Similarly, we can also partition $m_{i,t} = \Lambda_t \cdot \begin{pmatrix} m_{i,t-1} \\ \\ n_{i,t-1} \end{pmatrix} + \omega_{i,t}$ as the following form

$$\begin{pmatrix} \vec{m}_{i,t} \\ \\ \tilde{m}_{i,t} \end{pmatrix} = \begin{pmatrix} \Lambda_{a,t} & \Lambda_{b,t} \\ \\ \Lambda_{c,t} & \Lambda_{d,t} \end{pmatrix} \begin{pmatrix} \vec{s}_{i,t-1} \\ \\ x_{i,i,t|t-1} \end{pmatrix} + \begin{pmatrix} \vec{\omega}_{i,t} \\ \\ \tilde{\omega}_{i,t} \end{pmatrix} \tag{J.30}$$

Note that the elements of $\vec{s}_{i,t}$ are not correlated with $z_{i,i,t|t-1}$ and that $\vec{s}_{i,t-1}$ and $x_{i,i,t|t-1}$ span the same vector space as $(m_{i,t-1}, n_{i,t-1})$. I also impose the following normalization assumption

$$E[x_{i,i,t|t-1}\Big|\, m_{i,t}] = \tilde{m}_{i,t} + cons \cdot E[x_{i,i,t|t-1}]$$

This relationship holds if and only if $x_{i,i,t|t-1} - \tilde{m}_{i,t}$ is uncorrelated with all the elements in $m_{i,t}$. Two requirements summarize this relationship.

$$Cov\left[x_{i,i,t|t-1} - \tilde{m}_{i,t}, \vec{m}_{i,t}\right] = \vec{O} \tag{J.31a}$$

$$Cov\left[x_{i,i,t|t-1} - \tilde{m}_{i,t}, \tilde{m}_{i,t}\right] = O \tag{J.31b}$$

The second requirement implies that

$$\begin{aligned} & Cov\left[x_{i,i,t|t-1}, \tilde{m}_{i,t}\right] = V[\tilde{m}_{i,t}] \\ \Leftrightarrow \quad & V[\Lambda_{c,t}\,\vec{s}_{i,t-1} + \tilde{\omega}_{i,t}] = (1 - \Lambda_{d,t})\,V[x_{i,i,t|t-1}]\,\Lambda_{d,t}' \\ & \qquad = (1 - \Lambda_{d,t})\left(V[x_t] - \Sigma_{t|t-1}\right)\Lambda_{d,t}' \end{aligned}$$

The feasible set of $\Lambda_{d,t}$ is defined as the collection of $\Lambda_{d,t}$ under which the resulting right-hand side is a proper variance-covariance matrix (that is, symmetric and p.s.d.).

Step 2: Forecast accuracy depends only on $K_{d,t}$ and $\Lambda_{d,t}$

From the proposed partition (J.28), we can see that

$$x_t|\, m_{i,t}, n_{i,t} = x_t|\, m_{i,t}, \tilde{n}_{i,t}$$

That is, further knowledge of $\vec{n}_{i,t}$ does not improve the estimate of $x_t|\, m_{i,t}, \tilde{n}_{i,t}$. This follows from (J.29a). Furthermore, we can see that $K_{d,t}$ uniquely determines the posterior uncertainty $\Sigma_{t|t}$, given the prior uncertainty $\Sigma_{t|t}^m$,

Likewise, we can also see from (J.30) that

$$x_{i,i,t|t-1}\big|\, m_{i,t} = x_{i,i,t|t-1}\big|\, \tilde{m}_{i,t}$$

The information in $m_{i,t}$ about $x_{i,i,t|t-1}$ is completely captured by $\tilde{m}_{i,t}$, which follows from

(J.31a). We can furthermore see that $\Lambda_{d,t}$ uniquely determines the next-period prior uncertainty given $\Sigma_{t|t-1}$, the time-t posterior uncertainty about x_t.

$$\begin{aligned} V[x_t|\,\tilde{m}_t] &= V[x_t] - \Lambda_{d,t}\, Cov\left[x_t, x_{i,i,t|t-1}\right] \\ &= V[x_t] - \Lambda_{d,t}\left(V[x_t] - \Sigma_{x|t-1}\right) \\ &= (1-\Lambda_{d,t})\, V[x_t] + \Lambda_{d,t}\, \Sigma_{t|t-1} \end{aligned}$$

In summary, given $\Sigma_{t|t-1}$ at any time t, $K_{d,t}$ and $\Lambda_{d,t}$ uniquely determine $\Sigma^m_{t|t}$ and $\Sigma_{t|t}$. We can apply this argument recursively. It must be that the sequence of $\{K_{d,t}, \Lambda_{d,t}\}$ uniquely determines the sequence of $\left\{\Sigma_{t|t}\right\}$, given the initial prior uncertainty.

Step 3: The Optimal Choice of K_t and Λ_t

Since the remaining elements of K_t and Λ_t do not matter for the forecast accuracy, we can furthermore conclude that it is optimal to have them equal to zero. To see why note that

$$I\left(n_{i,t}; N_t\right) = I\left(\left(\vec{n}_{i,t}, \tilde{n}_{i,t}\right); \left(\vec{N}_t, E[x_t|\, N_t]\right)\right)$$

As discussed in Appendix C.2 of Azeredo da Silveira et al. (2020), the lower bound of this mutual information is equal to $I\left(\tilde{n}_{i,t}; E[x_t|\, N_t]\right)$. This lower bound is achieved when $K_{a,t} = K_{b,t} = K_{c,t} = O$. Likewise,

$$I\left(m_{i,t}; m_{i,t-1}, n_{i,t-1}\right) = I\left(\left(\vec{m}_{i,t}, \tilde{m}_{i,t}\right); \left(\vec{s}_{i,t-1}, x_{i,i,t|t-1}\right)\right)$$

whose lower bound is equal to $I\left(\tilde{m}_{i,t}; x_{i,i,t|t-1}\right)$. This lower bound is achieved when $\Lambda_{a,t} = \Lambda_{b,t} = \Lambda_{c,t} = O$.

10.2 Optimal Covnitive Process When z_t is the Only State Variable

In this section, I apply the result from the previous section when $x_t = z_t$.

Optimal representation of noisy news

The optimal $n_{i,t}$ is described as

$$\tilde{n}_{i,t} = \kappa_t \cdot E[\, z_t |\, N_t] + \tilde{u}_{i,t}$$

for some positive scalar κ_t. The idiosyncratic noise $\tilde{u}_{i,t}$ follows a Gaussian distribution $\mathcal{N}\left(0,\, \sigma^2_{u,t}\right)$, where $\sigma^2_{u,t}$ is determined by the choice of κ_t.

$$\sigma^2_{u,t} = \kappa_t \, Cov\left[\, z_t, E[\, z_t |\, N_t] |\, m_{i,t}\right] - \kappa_t^2 \, V[E[\, z_t |\, N_t] |\, m_{i,t}]$$

Without loss of generality, we could assume that $E[\, z_t |\, N_t]$ can be expressed as

$$E[\, z_t |\, N_t] = z_t + \bar{\nu}_t$$

where $\bar{\nu}_t \sim \mathcal{N}\left(0,\, \sigma^2_\nu\right)$ for some positive σ^2_ν. Then, $\sigma^2_{u,t}$ is further simplified to

$$\sigma^2_{u,t} = \kappa_t \left(1 - \kappa_t\right) \Sigma^m_{z,t|t} - \kappa_t^2 \, \sigma^2_\nu$$

where $\Sigma^m_{z,t|t} = V[\, z_t |\, m_{i,t}]$. Any $\kappa_t \in \left[0,\, \frac{\Sigma^m_{z,t|t}}{\Sigma^m_{z,t|t} + \sigma^2_\nu}\right]$ ensures that the resulting $\sigma^2_{u,t}$ is non-negative.

Determination of κ_t Using the information constraint, we can derive that

$$\begin{aligned} I\left(n_{i,t}; N_t\right) &= I\left(\tilde{n}_{i,t}; E[\,z_t|\,N_t]\right) \\ &= -\frac{1}{2}\log\left(1 - \frac{\kappa_t^2\, V[z_t + \tilde{\nu}_t]}{\kappa_t^2\, V[,\, z_t + \tilde{\nu}_t] + \sigma_{u,t}^2}\right) \\ &= -\frac{1}{2}\log\left(1 - \frac{V[z_t + \tilde{\nu}_t]}{V[z_t + \tilde{\nu}_t] + \left(\left(\kappa_t^{-1} - 1\right)\, \Sigma_{t|t}^m - \sigma_\nu^2\right)}\right) \leq -\frac{1}{2}\log\, \phi_n \end{aligned}$$

Rearranging the last inequality yields

$$\kappa_t \leq \frac{\Sigma_{t|t}^m}{\Sigma_{t|t}^m + \frac{\phi_n}{1-\phi_n}\left(V[z_t] + \sigma_\nu^2\right) + \sigma_\nu^2} \tag{J.32}$$

The upper bound is the optimal κ_t. Then, the resulting $\sigma_{u,t}^2$ is

$$\sigma_{u,t}^2 = \frac{\left(\Sigma_{t|t}^m\right)^2 \left(\frac{\phi_n}{1-\phi_n}\left(V[z_t] + \sigma_\nu^2\right)\right)}{\left(\Sigma_{t|t}^m + \frac{\phi_n}{1-\phi_n}\left(V[z_t] + \sigma_\nu^2\right) + \sigma_\nu^2\right)^2} \tag{J.33}$$

Optimal representation of noisy memory

Likewise, we can express the optimal $m_{i,t}$ as

$$\tilde{m}_{i,t} = \lambda_t \cdot z_{i,i,t|t-1} + \tilde{\omega}_{i,t}$$

for some positive scalar λ_t. The idiosyncratic noise $\tilde{\omega}_{i,t}$ follows a Gaussian distribution $\mathcal{N}\left(0,\, \sigma_{\omega,t}^2\right)$, whose variance is determined by the choice of λ_t as follows.

$$\sigma_{\omega,t}^2 = \lambda_t\left(1 - \lambda_t\right) V[z_{i,i,t|t-1}]$$

Any $\lambda_t \in [0, 1]$ ensures that the resulting $\sigma_{\omega,t}^2$ is non-negative.

Determination of λ_t Using the information constraint, we can derive that

$$\begin{aligned} I\left(m_{i,t}; m_{i,t-1}, n_{i,t-1}\right) &= I\left(\tilde{m}_{i,t}; z_{i,i,t|t-1}\right) \\ &= -\frac{1}{2}\log\det\left(1-\lambda_t\right) \leq -\frac{1}{2}\log\,\phi_m \end{aligned}$$

Therefore,

$$\lambda_t \leq 1-\phi_m$$

The optimal $\lambda_t = 1-\phi_m$ and the resulting $\sigma^2_{\omega,t} = \phi_m\left(1-\phi_m\right)\,V[z_{i,i,t|t-1}]$.

10.3 Optimal Cognitive Process When (μ, z_t) is the State Vector

In this section, I apply the result from Section 10.1 when $x_t = (\mu, z_t)$.

Optimal representation of noisy news

The optimal $n_{i,t}$ is described as

$$\tilde{n}_{i,t} = \tilde{K}_t \cdot E[x_t|\, N_t] + \tilde{u}_{i,t}$$

for some matrix scalar $\tilde{K}_t$. The idiosyncratic noise $\tilde{u}_{i,t}$ follows a Gaussian distribution $\mathcal{N}\left(O,\, \sigma_{u,t}\right)$, where $\sigma_{u,t}$ is determined by the choice of $\tilde{K}_t$.

$$\sigma_{u,t} = Cov\left[x_t, E[x_t|\, N_t]|\, m_{i,t}\right]\,\tilde{K}_t' - \tilde{K}_t\, V[E[x_t|\, N_t]|\, m_{i,t}]\,\tilde{K}_t'$$

Note that $E[x_t|\, N_t]$ is spanned by $E[z_t|\, N_t]$. This is because the news vector N_t is informative about μ only through the information about z_t. Therefore, without loss of generality,

we can express $\tilde{n}_{i,t}$ as

$$\tilde{n}_{i,t} = \begin{pmatrix} \frac{\kappa_{\mu,t}}{\kappa_t} \\ \\ 1 \end{pmatrix} \cdot \left(\kappa_t \, E[\, z_t |\, N_t] + \bar{u}_{i,t}\right)$$

where the idiosyncratic noise $\bar{u}_{i,t}$ is drawn from $\mathcal{N}\left(0,\, \sigma^2_{u,t}\right)$. The noisy news structure is then described by three univariate variables, $\kappa_{\mu,t}$, κ_t, and $\sigma^2_{u,t}$, which remain to be specified.

We could furthermore see that the normalization assumption $Cov\left[x_t, \tilde{n}_{i,t} |\, m_{i,t}\right] = V[\tilde{n}_{i,t} |\, m_{i,t}]$ implies that

$$\kappa_t \begin{pmatrix} \Sigma^m_{\mu,t|t}\left(\kappa_{\mu,t}/\kappa_t\right) & \Sigma^m_{\mu,t|t} \\ \\ \Sigma^m_{z,t|t}\left(\kappa_{\mu,t}/\kappa_t\right) & \Sigma^m_{z,t|t} \end{pmatrix} = \left(\kappa_t^2\left(\Sigma^m_{z,t|t} + \sigma^2_\nu\right) + \sigma^2_{u,t}\right) \begin{pmatrix} \left(\kappa_{\mu,t}/\kappa_t\right)^2 & \left(\kappa_{\mu,t}/\kappa_t\right) \\ \\ \left(\kappa_{\mu,t}/\kappa_t\right) & 1 \end{pmatrix}$$

where $\Sigma^m_{\mu,t|t} = V[\mu |\, m_{i,t}]$. This condition pins down $\kappa_{\mu,t}$ and $\sigma^2_{u,t}$ as a function of κ_t as follows.

$$\begin{aligned} \kappa_{\mu,t} &= \frac{\Sigma^m_{\mu,t|t}}{\Sigma^m_{z,t|t}}\, \kappa_t \\ \sigma^2_{u,t} &= \kappa_t\left(1 - \kappa_t\right) \Sigma^m_{z,t|t} - \kappa_t^2\, \sigma^2_\nu \end{aligned}$$

We can see that any $\kappa_t \in \left[0, \frac{\Sigma^m_{z,t|t}}{\Sigma^m_{z,t|t} + \sigma^2_\nu}\right]$ ensures a non-negative $\sigma^2_{u,t}$. Using $e = \begin{pmatrix} 0 & 1 \end{pmatrix}$ to pick out z_t from x_t, we have the following expression for $\tilde{n}_{i,t}$.

$$\tilde{n}_{i,t} = \tilde{K}_t \cdot E[x_t |\, N_t] + \tilde{u}_{i,t}, \quad \tilde{u}_{i,t} \sim \mathcal{N}\left(O,\, \sigma_{u,t}\right)$$

where $\tilde{K}_t$ and $\sigma_{u,t}$ are defined as

$$\tilde{K}_t = \kappa_t \frac{\Sigma^m_{t|t}\, e\, e'}{e'\, \Sigma^m_{t|t}\, e}$$

$$\sigma_{u,t} = \sigma^2_{u,t}\, \left(e'\, \Sigma^m_{t|t}\, e\right)^{-2} \Sigma^m_{t|t}\, e\, e'\, \Sigma^m_{t|t}$$

for which we use the relationship $\begin{pmatrix} \frac{\kappa_{\mu,t}}{\kappa_t} \\ \\ 1 \end{pmatrix} = \frac{\Sigma^m_{t|t}\, e}{e'\, \Sigma^m_{t|t}\, e}$.

Determination of κ_t We can observe that the optimal κ_t and $\sigma^2_{u,t}$ are equal to the ones determined in Section 10.2. This is because the optimal $n_{i,t}$ under the state vector $x_t = (\mu, z_t)$ is spanned from the optimal $n_{i,t}$ when $x_t = z_t$. The information constraint (B.4) has the same restriction.

Posterior beliefs It is straightforward to see that the posterior belief evolves as follows, given $\tilde{n}_{i,t}$.

$$x_{i,i,t|t} = \left(I - \tilde{K}_t\right) x^m_{i,i,t|t} + \tilde{K}_t\, x_t + \tilde{\nu}_t + \tilde{u}_{i,t}$$

$$\Sigma_{t|t} = \left(I - \tilde{K}_t\right) \Sigma^m_{t|t}$$

where $\tilde{\nu}_t \sim \mathcal{N}\left(O,\, \sigma_\nu\right)$ and $\sigma_\nu = \kappa^2_t\, \left(e'\, \Sigma^m_{t|t}\, e\right)^{-2} \Sigma^m_{t|t}\, e\, e'\, \Sigma^m_{t|t}$.

Optimal representation of noisy memory

We can express the optimal $m_{i,t}$ as

$$\tilde{m}_{i,t} = \tilde{\Lambda}_t \cdot x_{i,i,t|t-1} + \tilde{\omega}_{i,t}$$

The feasibility of $\tilde{\Lambda}_t$ is described earlier. The idiosyncratic noise $\tilde{\omega}_{i,t}$ follows a Gaussian distribution $\mathcal{N}(O,\, \sigma_{\omega,t})$, whose variance is determined by the choice of $\tilde{\Lambda}_t$ as follows.

$$\sigma_{\omega,t} = \left(1 - \tilde{\Lambda}_t\right) \left(V[x_t] - \Sigma_{t|t-1}\right) \tilde{\Lambda}_t'$$

Therefore, it remains to specify $\tilde{\Lambda}_t$. The information constraint (B.6) constrains the choice of $\tilde{\Lambda}_t$. We can derive that

$$\begin{aligned} I\left(m_{i,t}; m_{i,t-1}, n_{i,t-1}\right) &= I\left(\tilde{m}_{i,t}; x_{i,i,t|t-1}\right) = h\left(\tilde{m}_{i,t}\right) - h\left(\tilde{m}_{i,t} |\, x_{i,i,t|t-1}\right) \\ &= \frac{1}{2} \ln \det\left(V[\tilde{m}_{i,t}]\right) - \frac{1}{2} \ln \det\left(V[\tilde{m}_{i,t} |\, x_{i,i,t|t-1}]\right) \\ &= \frac{1}{2} \ln \det\left(V[x_{i,i,t|t-1}]\, \tilde{\Lambda}_t'\right) - \frac{1}{2} \ln \det\left(\left(I - \tilde{\Lambda}_t\right) V[x_{i,i,t|t-1}]\, \tilde{\Lambda}_t'\right) \\ &= -\frac{1}{2} \log \det\left(1 - \tilde{\Lambda}_t\right) \leq -\frac{1}{2} \log\, \phi_m \end{aligned}$$

Therefore,

$$\det\left(I - \tilde{\Lambda}_t\right) \geq \phi_m$$

The Choice Variable

Any $\tilde{\Lambda}_t$ is feasible as long as (1) the resulting $\Sigma_{t|t}^m$ is a symmetric and positive semidefinite matrix and (2) the diagonal elements of $\Sigma_{t|t}^m$ are bigger than those of $\Sigma_{t|t-1}$ and smaller than those of σ_x. That is, under any feasible $\tilde{\Lambda}_t$, both $\Sigma_{t|t}^m - \Sigma_{t|t-1}$ and $\sigma_x - \Sigma_{t|t}^m$ are proper variance-covariance matrices (symmetric and positive semidefinite).

It is useful to define $\bar{\Lambda}_t$, which is simply a rotation of $\tilde{\Lambda}_t$.

$$\bar{\Lambda}_t = V[x_{i,i,t|t-1}]^{-\frac{1}{2}} \tilde{\Lambda}_t\, V[x_{i,i,t|t-1}]^{\frac{1}{2}}$$

We could confirm that the same accuracy constraint (B.6) applies.

$$\begin{aligned}\det\left(I-\bar{\Lambda}_t\right) &= \det\left(I - V[x_{i,i,t|t-1}]^{-\frac{1}{2}}\tilde{\Lambda}_t V[x_{i,i,t|t-1}]^{\frac{1}{2}}\right) \\ &= \det\left(V[x_{i,i,t|t-1}]^{-\frac{1}{2}}\left(I-\tilde{\Lambda}_t\right)V[x_{i,i,t|t-1}]^{\frac{1}{2}}\right) = \det\left(I-\tilde{\Lambda}_t\right)\end{aligned}$$

Therefore, I use $W_t = I - \bar{\Lambda}_t$ as a choice variable. Any W_t is feasible as long as W_t and $I - W_t$ are positive semidefinite.

The Constraints

The prior uncertainty is formed according to

$$\begin{aligned}\Sigma^m_{t|t} &= \Sigma_{t|t-1} + \left(I-\tilde{\Lambda}_t\right)V[x_{i,i,t|t-1}] \\ &= \Sigma_{t|t-1} + V[x_{i,i,t|t-1}]^{\frac{1}{2}}\left(I-\bar{\Lambda}_t\right)V[x_{i,i,t|t-1}]^{\frac{1}{2}}\end{aligned}$$

And the posterior uncertainty can be described as

$$\begin{aligned}\Sigma_{t|t} &= \Sigma^m_{t|t} - \left(\kappa_t\,\Sigma^m_{t|t}\,e\right)\left(\kappa_t\,e'\,\Sigma^m_{t|t}\,e\right)^{-1}\left(\kappa\,e'\,\Sigma^m_{t|t}\right) \\ &= \Sigma^m_{t|t} - \Sigma^m_{t|t}\,e\,\left(\Omega^m_{t|t}\right)^{-1}e'\,\Sigma^m_{t|t}\end{aligned}$$

where

$$\Omega^m_{t|t} = e'\,\Sigma^m_{t|t}\,e + \frac{\phi_n}{1-\phi_n}\left(V[z_t]+\sigma_\nu\right)+\sigma^2_\nu$$

The Optimization Problem

The optimization problem can then be written as

$$\min_{W_t} tr\left(\sigma_{t|t}\,Q\right)$$

subject to the law of motions of the subjective uncertainty

$$\begin{aligned}
\Sigma^m_{t|t} - \Sigma_{t|t-1} &= \left(\sigma_x - \Sigma_{t|t-1}\right)^{\frac{1}{2}} W_t \left(\sigma_x - \Sigma_{t|t-1}\right)^{\frac{1}{2}} \\
\Omega^m_{t|t} &= e' \Sigma^m_{t|t} e + \frac{\phi_n}{1-\phi_n} \left(V[z_t] + \sigma_\nu\right) + \sigma^2_\nu \\
\sigma_{t|t} &= \Sigma^m_{t|t} - \Sigma^m_{t|t} e \left(\Omega^m_{t|t}\right)^{-1} e' \Sigma^m_{t|t}
\end{aligned}$$

along with the requirement that both W_t and $I - W_t$ are positive semidefinite and symmetric.

Note that when deciding which information to recall at time t (or equivalently, when deciding which information to store at time $t-1$), such a decision takes into account the noisy news that is available at time t. That is, the availability (and the quality) of extra information not from one's memory will affect which information is worthy of remembering. While this is a natural trade-off given the restriction that memory cannot perfectly store all the past information, it is also one that has not been investigated in the literature yet.

Setting up the Lagrange Multipliers

Since W_t is symmetric, it can be eigen-decomposed as $W_t = U\ (I - D)\ U'$ where D is a diagonal matrix and $U\,U' = I$. The constraints that W_t and $I - W_t$ are positive semidefinite are equivalent to the constraints that $I - D$ and D are positive semidefinite. The diagonal elements of $I - D$ and D should be non-negative. The Lagrange multipliers for each inequality constraint can be stored in a diagonal matrix, $\bar{\Upsilon}_1$ and $\bar{\Upsilon}_1$. Finally, I can define $\Upsilon_1 = U\,\bar{\Upsilon}_1\,U'$ and $\Upsilon_2 = U\,\bar{\Upsilon}_2\,U'$. Note that $\Upsilon_1\,W_t = U\,\bar{\Upsilon}_1\,(I - D)\ U'$ and $\Upsilon_2\ (I - W_t) = U\,\bar{\Upsilon}_2\,(D)\ U'$. We can see that the inequality constraint can be expressed as $tr\left(\Upsilon_1\,W_t\right) \geq 0$ and $tr\left(\Upsilon_2\ (I - W_t)\right) \geq 0$. This is because $tr\left(\Upsilon_1\,W_t\right) = tr\left(\bar{\Upsilon}_1\,(I - D)\right)$ and $tr\left(\Upsilon_2\ (I - W_t)\right) = tr\left(\bar{\Upsilon}_2\,(D)\right)$.

We also have equality constraints on the law of motions of subjective uncertainty. For each constraint, I construct a symmetric matrix Γ_i whose kth row contains the Lagrangian

multiplier for each kth column of the equality conditions.

The Lagrangian Problem and the First Order Conditions

The Lagrangian problem is as follows.

$$\begin{aligned} max & - tr\left(\sigma_{t|t}\, Q\right) \\ & - tr\left(\Gamma_1 \left(\left(\sigma_x - \Sigma_{t|t-1}\right)^{\frac{1}{2}} W_t \left(\sigma_x - \Sigma_{t|t-1}\right)^{\frac{1}{2}} + \Sigma_{t|t-1} - \Sigma^m_{t|t}\right)\right) \\ & - tr\left(\Gamma_2 \left(e'\, \Sigma^m_{t|t}\, e + \frac{\phi_n}{1-\phi_n}\left(V[z_t] + \sigma_\nu\right) + \sigma^2_\nu - \Omega^m_{t|t}\right)\right) \\ & - tr\left(\Gamma_3 \left(\Sigma^m_{t|t} - \Sigma^m_{t|t}\, e\, \left(\Omega^m_{t|t}\right)^{-1} e'\, \Sigma^m_{t|t} - \sigma_{t|t}\right)\right) \\ & + tr\left(\Upsilon_1 W_t\right) + tr\left(\Upsilon_2 \left(I - W_t\right)\right) + \mu\, \left(\det\left(W_t\right) - \phi_m\right) \end{aligned}$$

where the "Langrangian multipliers" Γ_i and Υ_i for all i are symmetric matrices.

The first order conditions subject to W_t, $\Sigma^m_{t|t}$, $\Omega^m_{t|t}$ and $\sigma_{t|t}$ are (in that order)

$$-\left(\sigma_x - \Sigma_{t|t-1}\right)^{\frac{1}{2}} \Gamma_1 \left(\sigma_x - \Sigma_{t|t-1}\right)^{\frac{1}{2}} + \Upsilon_1 - \Upsilon_2 + \mu\, \det\left(W_t\right) W_t^{-1} = O \tag{J.34a}$$

$$\Gamma_1 - e\, \Gamma_2\, e' - \Gamma_3 + e\left(\Omega^m_{t|t}\right)^{-1} e'\, \Sigma^m_{t|t}\, \Gamma_3 + \Gamma_3\, \Sigma^m_{t|t}\, e\left(\Omega^m_{t|t}\right)^{-1} e' = O \tag{J.34b}$$

$$\Gamma_2 - \left(\Omega^m_{t|t}\right)^{-1} e'\, \Sigma^m_{t|t}\, \Gamma_3\, \Sigma^m_{t|t}\, e\, \left(\Omega^m_{t|t}\right)^{-1} = O \tag{J.34c}$$

$$-Q + \Gamma_3 = O \tag{J.34d}$$

and the slackness conditions are

$$\Upsilon_1 W_t = O,\ \Upsilon_1 \succeq O,\ W_t \succeq O \tag{J.35a}$$

$$\Upsilon_2 \left(I - W_t\right) = O,\ \Upsilon_2 \succeq O,\ \left(I - W_t\right) \succeq O \tag{J.35b}$$

and

$$\mu\, \left(\det\left(W_t\right) - \phi_m\right) = 0,\ \mu \geq 0, \det\left(W_t\right) = \phi_m \tag{J.36}$$

We can first rearrange (J.34b)-(J.34d). Note that $\Gamma_3 = Q$ (as implied by (J.34d)) and using the notation $\tilde{K}_t \equiv \Sigma^m_{t|t}\, e\, \left(\Omega^m_{t|t}\right)^{-1} e'$, we can express (J.34b) as

$$\Gamma_1 - e\,\Gamma_2\,e' - Q + \tilde{K}_t'\,Q + Q\,\tilde{K}_t = O$$

and (J.34c) as

$$e\,\Gamma_2\,e' - \tilde{K}_t'\,Q\,\tilde{K}_t = O$$

which together result in

$$\Gamma_1 = \left(I - \tilde{K}_t\right)'\,Q\,\left(I - \tilde{K}_t\right)$$

Next, I'd like to solve for W_t that characterizes the optimal memory system. First, multiplying (J.34a) by $W_t\ (I - W_t)$ on the left yields

$$-\left(\sigma_x - \Sigma_{t|t-1}\right)^{\frac{1}{2}}\Gamma_1\left(\sigma_x - \Sigma_{t|t-1}\right)^{\frac{1}{2}}\ W_t\ (I - W_t) + \mu\,\phi_m\ (I - W_t) = O \tag{J.37}$$

after applying the slackness conditions (from which $(\Upsilon_1 - \Upsilon_2)\,W_t\ (I - W_t) = O$). We can observe that $\left(\sigma_x - \Sigma_{t|t-1}\right)^{\frac{1}{2}}\Gamma_1\left(\sigma_x - \Sigma_{t|t-1}\right)^{\frac{1}{2}}$ should be eigen-decomposed in the form of $U\,G\,U'$, that is, it should share the basis with Υ_1, Υ_2 and W_t. Then, the above expression can be written as

$$U\ (\mu\,\phi_m\,I - G\ (I - D))\,D\,U' = O \tag{J.38}$$

Note that D should satisfy $D \succeq O$, $I - D \succeq O$, and $\det(I - D) = \phi_m$.

The Solution to the Lagrangian Problem

The solution of D can be found as follows. Let's first rearrange U and G so that the diagonal elements in G are in descending order. For $k = 1, \cdots, n$ (where n is the dimension of x_t), I define $\theta_k = \left(\phi_m \prod_{i=1}^{k} g_i\right)^{\frac{1}{k}}$ then we can find k such that $g_k \geq \theta_k > g_{k+1}$ for $k < n$ (or $k = n$ if $g_k \geq \theta_k$). Then, the ith element of D, d_i, is going to be

$$d_i = \begin{cases} 1 - \frac{\theta_k}{g_i} & \text{for } i \leq k \\ 0 & \text{for } i > k \end{cases}$$

We can see that all $d_i \in [0, 1]$ and $\det(I - D) = \Pi_{i=1}^{k} \frac{\theta_k}{g_i} = \phi_m$.

We can express the solution for D more succinctly. Following Afrouzi and Yang (2021), I adopt the following two matrix operators. For a diagonal matrix D, $\max(D, \theta)$ replaces the diagonal elements of D that are smaller than θ with θ. For a symmetric matrix X whose eigendecomposition is expressed as $X = U\,D\,U'$, the operator $Max(X, \theta)$ is defined as $Max(X, \theta) = U \max(D, \theta) U'$. Using these operators, I can express the optimal $I - D$ as

$$I - D = \theta_k \{Max(G, \theta_k)\}^{-1}$$

Since $W_t = U(I - D)U'$, the optimal solution for W_t is expressed as

$$W_t = \theta_k \{Max(U\,G\,U', \theta_k)\}^{-1}$$

From this, the optimal $\Sigma_{t|t}^{m}$ is derived as

$$\Sigma_{t|t}^{m} = \Sigma_{t|t-1} + V[x_{i,i,t|t-1}]^{\frac{1}{2}}\, \theta_k \left\{Max\left(V[x_{i,i,t|t-1}]^{\frac{1}{2}}\, \Gamma_1\, V[x_{i,i,t|t-1}]^{\frac{1}{2}}, \theta_k\right)\right\}^{-1} V[x_{i,i,t|t-1}]^{\frac{1}{2}}$$

where $V[x_{i,i,t|t-1}] = \sigma_X - \Sigma_{t|t-1}$ captures the maximum possible increase in the uncertainty due to forgetting the previous information $s_{i,t-1}$. In summary, the optimal memory system

solves the fixed point problem for Γ_1 and $\Sigma^m_{t|t}$ that satisfy the following equations, given the level of $\Sigma_{t|t-1}$ (and therefore $V[x_{i,i,t|t-1}]$).

$$\Sigma^m_{t|t} = \Sigma_{t|t-1} + V[x_{i,i,t|t-1}]^{\frac{1}{2}}\,\theta_k \left\{ Max\left(V[x_{i,i,t|t-1}]^{\frac{1}{2}}\,\Gamma_1\, V[x_{i,i,t|t-1}]^{\frac{1}{2}}, \theta_k \right) \right\}^{-1} V[x_{i,i,t|t-1}]^{\frac{1}{2}}$$
$$\Gamma_1 = \left(I - \tilde{K}_t\right)' Q \left(I - \tilde{K}_t\right)$$

Furthermore, as summarized by $\tilde{\Lambda}_t$, the optimal memory signal is described as follows.

$$\tilde{\Lambda}_t = V[x_{i,i,t|t-1}]^{\frac{1}{2}} \left(\sum_{i=1}^{k} \left(1 - \frac{\theta_k}{g_i}\right) u_i\, u_i' \right) V[x_{i,i,t|t-1}]^{-\frac{1}{2}}$$

where g_i is the eigenvalues of $V[x_{i,i,t|t-1}]^{\frac{1}{2}}\,\Gamma_1\, V[x_{i,i,t|t-1}]^{\frac{1}{2}}$ that are rearranged in a descending order and u_i is the corresponding eigenvector. As defined above, k is such that $g_k \geq \theta_k \geq g_{k+1}$.

10.4 Derivations of β_I and β_C (when the long-run mean is known)

DM i's forecast of z_t evolves according to the following linear law of motion.

$$z_{i,t|t} = (1-\lambda)(1-\kappa)\mu + \lambda\,(1-\kappa)\,z_{i,t|t-1} + \kappa\,z_t + \kappa\,\tilde{\nu}_t + \tilde{u}_{i,t} + (1-\kappa)\,\tilde{\omega}_{i,t}$$

The consensus forecast of z_t evolves according to the following linear law of motion.

$$z_{t|t} = (1-\lambda)(1-\kappa)\mu + \lambda\,(1-\kappa)\,z_{t|t-1} + \kappa\,z_t + \kappa\,\tilde{\nu}_t \tag{J.39}$$

I define b as the weight on unconditional prior belief.

$$b \equiv (1-\lambda)(1-\kappa) \tag{J.40}$$

Derivations of β_I and β_C

Derivation of β_I

From the regression specification

$$z_t - z_{i,t|t} = \alpha_I + \beta_I\left(z_{i,t|t} - z_{i,t|t-1}\right) + error_{i,t},$$

the coefficient β_I asymptotically converges to

$$\beta_I = \frac{Cov\left[z_t - z_{i,t|t},\, z_{i,t|t} - z_{i,t|t-1}\right]}{V[z_{i,t|t} - z_{i,t|t-1}]}$$

We can see that

$$Cov\left[z_t - z_{i,t|t},\, z_{i,t|t} - z_{i,t|t-1}\right] = -Cov\left[z_t - z_{i,t|t},\, z_{i,t|t-1}\right] = -b\,V[z_{i,t|t}]$$

The first equality holds because $Cov\left[z_t - z_{i,t|t}, z_{i,t|t}\right] = 0$. The second equality holds because $E[\, z_{i,t|t} \big|\, m_{i,t-1}, n_{i,t-1}] = b\,\mu + (1-b)\, z_{i,t|t-1}$. We can also see that

$$V[z_{i,t|t} - z_{i,t|t-1}] = \left(\rho^{-2} - 2\,(1-b) + 1\right) V[z_{i,t|t-1}]$$

where I use $V[z_{i,t|t-1}] = \rho^2\, V[z_{i,t|t}]$. Combining the two derivations, we get

$$\beta_I = -\frac{b}{2\,b + \rho^{-2} - 1} \tag{J.41}$$

Derivation of β_C

Rearranging terms, we can express the consensus forecast's error as follows.

$$z_t - z_{t|t} = \frac{1-\kappa}{\kappa}\left(z_{t|t} - z_{t|t-1} + (1-\lambda)\left(z_{t|t-1} - \mu\right)\right) - \tilde{\nu}_t$$

From the regression specification

$$z_t - z_{t|t} = \alpha_C + \beta_C\left(z_{t|t} - z_{t|t-1}\right) + error_t,$$

the coefficient β_C asymptotically converges to

$$\beta_C = \frac{Cov\left[z_t - z_{t|t}, z_{t|t} - z_{t|t-1}\right]}{V[z_{t|t} - z_{t|t-1}]}$$

Therefore, we can see that

$$\beta_C = \frac{1-\kappa}{\kappa}\left(1 + (1-\lambda)\,\frac{Cov\left[z_{t|t-1}, z_{t|t} - z_{t|t-1}\right]}{V[z_{t|t} - z_{t|t-1}]}\right) - \frac{\kappa\,\sigma_\nu^2}{V[z_{t|t} - z_{t|t-1}]}$$

It remains to derive expressions for $Cov\left[z_{t|t-1}, z_{t|t} - z_{t|t-1}\right]$ and $V[z_{t|t} - z_{t|t-1}]$.

Note that

$$\begin{aligned}(1-\lambda(1-\kappa)\rho L)\, z_{t|t} &= \kappa(z_t+\tilde{\nu}_t)\\ \Leftrightarrow \quad z_{t|t} &= \frac{\kappa}{1-\lambda(1-\kappa)\rho L}(z_t+\tilde{\nu}_t)\end{aligned}$$

Therefore, it is straightforward to see that

$$Cov\left[z_t, z_{t|t}\right] = \frac{\kappa}{1-\lambda(1-\kappa)\rho^2} V[z_t]$$

We can also show that

$$\begin{aligned}V[z_{t|t}] &= V[\frac{\kappa}{1-\lambda(1-\kappa)\rho L}\frac{1}{1-\rho L}\epsilon_t + \frac{\kappa}{1-\lambda(1-\kappa)\rho L}\tilde{\nu}_t]\\ &= \left[\frac{1+\lambda(1-\kappa)\rho^2}{1-\lambda(1-\kappa)\rho^2}\frac{\kappa^2}{1-(\lambda(1-\kappa)\rho)^2}\frac{\sigma_\epsilon^2}{1-\rho^2}\right] + \left[\frac{\kappa^2}{1-(\lambda(1-\kappa)\rho)^2}\sigma_\nu^2\right]\\ &= \frac{\kappa^2}{1-(\lambda(1-\kappa)\rho)^2}\left\{\frac{1+\lambda(1-\kappa)\rho^2}{1-\lambda(1-\kappa)\rho^2}V[z_t] + \sigma_\nu^2\right\}\end{aligned}$$

And finally,

$$Cov\left[z_{t|t}, z_{t|t-1}\right] = \lambda(1-\kappa)\rho^2 V[z_{t|t}] + \kappa\rho^2 Cov\left[z_t, z_{t|t}\right]$$

Let's consider the case $\sigma_\nu^2 \to 0$. Then,

$$\begin{aligned}Cov\left[z_t, z_{t|t}\right] &= \frac{1}{k}\frac{1-(\lambda(1-\kappa)\rho)^2}{1+\lambda(1-\kappa)\rho^2}V[z_{t|t}]\\ Cov\left[z_{t|t}, z_{t|t-1}\right] &= \left[\kappa\rho^2 + \kappa\rho^2\frac{1}{k}\frac{1-(\lambda(1-\kappa)\rho)^2}{1+\lambda(1-\kappa)\rho^2}\right]V[z_{t|t}]\\ &= \frac{\rho^2+\lambda(1-\kappa)\rho^2}{1+\lambda(1-\kappa)\rho^2}V[z_{t|t}] \equiv \bar{c}\,V[z_{t|t}]\end{aligned}$$

Then,

$$\frac{Cov\left[z_{t|t-1}, z_{t|t} - z_{t|t-1}\right]}{V[z_{t|t} - z_{t|t-1}]} = \frac{(\bar{c} - \rho^2)\ V[z_{t|t}]}{(1 + \rho^2 - 2\,\bar{c})\, V[z_{t|t}]} = \frac{\bar{c} - \rho^2}{1 + \rho^2 - 2\,\bar{c}} = \frac{\lambda\,(1-\kappa)\,\rho^2}{1 - \lambda\,(1-\kappa)\,\rho^2}$$

Finally, we can derive that β_C is expressed as follows.

$$\beta_C = \frac{1-\kappa}{\kappa}\left(1 + (1-\lambda)\,\frac{\lambda\,(1-\kappa)\,\rho^2}{1 - \lambda\,(1-\kappa)\,\rho^2}\right) \tag{J.42}$$

Steady-state Uncertainty

I denote the steady state uncertainty of z_t as $\Sigma_{-1} \equiv V[\,z_t|\,m_{i,t-1}, n_{i,t-1}]$, $\Sigma^m \equiv V[\,z_t|\,m_{i,t}]$, and $\Sigma \equiv V[\,z_t|\,m_{i,t}, n_{i,t}]$, which satisfy the following stationary relationship.

$$\Sigma_{-1} = \rho^2\,\Sigma + \sigma_\epsilon^2 \tag{J.43a}$$

$$\Sigma^m = (1-\lambda)\,\sigma_z^2 + \lambda\,\Sigma_{-1} \tag{J.43b}$$

$$(\Sigma)^{-1} = (\Sigma^m)^{-1} + \left(\tilde{\sigma}_u^2\right)^{-1} \tag{J.43c}$$

where σ_z^2 is the unconditional variance of z, which equals $= \frac{\sigma_\epsilon^2}{1-\rho^2}$, and $\tilde{\sigma}_u^2 = \frac{\phi_n}{1-\phi_n}\,\sigma_z^2$ captures the noisy news.

The steady-state κ and b are

$$\kappa = \frac{\Sigma^m}{\Sigma^m + \tilde{\sigma}_u^2} \tag{J.44}$$

$$b = (1-\lambda)\ \frac{\tilde{\sigma}_u^2}{\Sigma^m + \tilde{\sigma}_u^2} \tag{J.45}$$

And we have shown earlier that $\lambda = 1 - \phi_m$.

Comparative Statics

Comparative Statics for the Uncertainty

Equations (J.43) implicitly impose the following relation.

$$\begin{aligned} F\left(\Sigma;\tilde{\sigma}_u^2,\lambda\right) &= \left(\Sigma^m\right)^{-1} + \left(\tilde{\sigma}_u^2\right)^{-1} - \left(\Sigma\right)^{-1} \\ &= \left(\left(1-\lambda\right)\sigma_z^2 + \lambda\,\left(\rho^2\,\Sigma + \sigma_\epsilon^2\right)\right)^{-1} + \left(\tilde{\sigma}_u^2\right)^{-1} - \left(\Sigma\right)^{-1} = 0 \end{aligned} \tag{J.46}$$

Then, the derivatives of $F\left(\Sigma;\tilde{\sigma}_u^2,\lambda\right)=0$ with respect to $\tilde{\sigma}_u^2$ and λ are

$$\begin{aligned} \frac{\partial F}{\partial \tilde{\sigma}_u^2} &= -\left(\Sigma^m\right)^{-2}\lambda\,\rho^2\,\frac{\partial \Sigma}{\partial \tilde{\sigma}_u^2} - \left(\tilde{\sigma}_u^2\right)^{-2} + \left(\Sigma\right)^{-2}\frac{\partial \Sigma}{\partial \tilde{\sigma}_u^2} = 0 \\ \frac{\partial F}{\partial \lambda} &= -\left(\Sigma^m\right)^{-2}\left(-\sigma_z^2 + \rho^2\Sigma + \sigma_\epsilon^2 + \lambda\,\rho^2\,\frac{\partial \Sigma}{\partial \lambda}\right) + \left(\Sigma\right)^{-2}\frac{\partial \Sigma}{\partial \lambda} = 0 \end{aligned}$$

Rearranging yields the derivatives of σ with respect to $\tilde{\sigma}_u^2$ and λ.

$$\begin{aligned} \frac{\partial \Sigma}{\partial \tilde{\sigma}_u^2} &= \left(\left(\frac{\Sigma^m}{\Sigma}\right)^2 - \lambda\,\rho^2\right)^{-1}\left(\frac{\Sigma^m}{\tilde{\sigma}_u^2}\right)^2 > 0 \\ \frac{\partial \Sigma}{\partial \lambda} &= -\left(\left(\frac{\Sigma^m}{\Sigma}\right)^2 - \lambda\,\rho^2\right)^{-1}\left(\sigma_z^2 - \Sigma_{-1}\right) \\ &= -\left(\left(\frac{\Sigma^m}{\sigma}\right)^2 - \lambda\,\rho^2\right)^{-1}\frac{\Sigma^m}{1-\lambda}\left(1 - \frac{\Sigma_{-1}}{\Sigma^m}\right) < 0 \end{aligned}$$

Additionally, the derivative of Σ^m with respect to $\tilde{\sigma}_u^2$ is

$$\frac{\partial \Sigma^m}{\partial \tilde{\sigma}_u^2} = \lambda\,\rho^2\,\frac{\partial \Sigma}{\partial \tilde{\sigma}_u^2} = \lambda\,\rho^2\left(\left(\frac{\Sigma^m}{\Sigma}\right)^2 - \lambda\,\rho^2\right)^{-1}\left(\frac{\Sigma^m}{\tilde{\sigma}_u^2}\right)^2 > 0$$

and with respect to λ:

$$\begin{aligned}\frac{\partial \Sigma^m}{\partial \lambda} &= -\rho^2\left(\sigma_z^2 - \Sigma\right) + \lambda\,\rho^2\,\frac{\partial \Sigma}{\partial \lambda} \\ &= -\frac{\Sigma^m}{1-\lambda}\left(1 - \frac{\Sigma_{-1}}{\Sigma^m}\right)\left\{1 + \frac{\lambda\,\rho^2}{\left(\frac{\Sigma^m}{\Sigma}\right)^2 - \lambda\rho^2}\right\} \\ &= -\frac{\Sigma^m}{1-\lambda}\frac{1 - \frac{\Sigma_{-1}}{\Sigma^m}}{1 - \lambda\rho^2\,\left(\frac{\Sigma}{\Sigma^m}\right)^2} < 0\end{aligned}$$

Note that $1 > \frac{\Sigma_{-1}}{\Sigma^m} > \frac{\Sigma}{\Sigma^m} > \lambda\rho^2\,\left(\frac{\Sigma}{\Sigma^m}\right)^2 > 0$, making the last term be between 0 and 1.

Comparative Statics for κ and b

Now we turn to the comparative statistics of κ and b. First, the derivative of b with respect to $\tilde{\sigma}_u^2$ is computed as:

$$\begin{aligned}\frac{\partial b}{\partial \tilde{\sigma}_u^2} &= (1-\lambda)\,\frac{1}{\left(\Sigma^m + \tilde{\sigma}_u^2\right)^2}\left\{\left(\Sigma^m + \tilde{\sigma}_u^2\right) - \tilde{\sigma}_u^2\,\left(\frac{\partial \Sigma^m}{\partial \tilde{\sigma}_u^2} + 1\right)\right\} \\ &= (1-\lambda)\,\frac{\Sigma^m}{\left(\Sigma^m + \tilde{\sigma}_u^2\right)^2}\left\{1 - \lambda\,\rho^2\,\frac{\frac{\Sigma^m}{\Sigma} - 1}{\left(\frac{\Sigma^m}{\Sigma}\right)^2 - \lambda\rho^2}\right\} > 0\end{aligned}$$

We can easily see that $\frac{\frac{\Sigma^m}{\Sigma} - 1}{\left(\frac{\Sigma^m}{\Sigma}\right)^2 - \lambda\rho^2} \in (0,1)$, which makes the term inside the bracket be positive. Next, the derivative of b with respect to λ is derived as:

$$\begin{aligned}\frac{\partial b}{\partial \lambda} &= -\frac{\tilde{\sigma}_u^2}{\Sigma_m + \tilde{\sigma}_u^2} - (1-\lambda)\,\frac{\tilde{\sigma}_u^2}{\left(\Sigma_m + \tilde{\sigma}_u^2\right)^2}\frac{\partial \Sigma^m}{\partial \lambda} = -\frac{\tilde{\sigma}_u^2}{\Sigma_m + \tilde{\sigma}_u^2}\left(1 + \frac{1-\lambda}{\Sigma_m + \tilde{\sigma}_u^2}\frac{\partial \Sigma^m}{\partial \lambda}\right) \\ &= -\frac{\tilde{\sigma}_u^2}{\Sigma_m + \tilde{\sigma}_u^2}\left(1 - \frac{\Sigma^m}{\Sigma_m + \tilde{\sigma}_u^2}\frac{1 - \frac{\Sigma_{-1}}{\Sigma^m}}{1 - \lambda\rho^2\,\left(\frac{\Sigma}{\Sigma^m}\right)^2}\right) < 0\end{aligned}$$

In addition, the derivative of κ with respect to $\tilde{\sigma}_u^2$ is:

$$\begin{aligned}\frac{\partial \kappa}{\partial \tilde{\sigma}_u^2} &= -\frac{\Sigma^m}{\left(\Sigma^m + \tilde{\sigma}_u^2\right)^2}\left\{1 - \frac{\partial \Sigma^m}{\partial \tilde{\sigma}_u^2}\frac{\tilde{\sigma}_u^2}{\Sigma^m}\right\} \\ &= -\frac{\Sigma^m}{\left(\Sigma^m + \tilde{\sigma}_u^2\right)^2}\left\{1 - \lambda\rho^2\frac{\frac{\Sigma^m}{\Sigma} - 1}{\left(\frac{\Sigma^m}{\Sigma}\right)^2 - \lambda\rho^2}\right\} < 0\end{aligned}$$

Finally, the derivative of κ with respect to λ:

$$\frac{\partial \kappa}{\partial \lambda} = \frac{\tilde{\sigma}_u^2}{\left(\Sigma^m + \tilde{\sigma}_u^2\right)^2}\frac{\partial \Sigma^m}{\partial \lambda} < 0$$

Comparative Statics for β_I

Now we combine the above comparative statistics to analyze how β_I and β_C change with ϕ_n and ϕ_m. Note first from (J.41) that ϕ_n and ϕ_m affect β_I through the bias term b. The derivative of β_I with respect to b is:

$$\frac{\partial \beta_I}{\partial b} = -\left(2\,b + \rho^{-2} - 1\right)^{-2}\left(\rho^2 - 1\right) < 0$$

Therefore, we get that

$$\frac{\partial \beta_I}{\partial \phi_m} = \frac{\partial \beta_I}{\partial b}\frac{\partial b}{\partial \phi_m} = -\frac{\partial \beta_I}{\partial b}\frac{\partial \beta_I}{\partial \lambda} < 0 \tag{J.47a}$$

$$\frac{\partial \beta_I}{\partial \phi_n} = \frac{\partial \beta_I}{\partial b}\frac{\partial b}{\partial \tilde{\sigma}_u^2}\frac{\partial \tilde{\sigma}_u^2}{\partial \phi_n} < 0 \tag{J.47b}$$

Comparative Statics for β_C

Next, we analyze the comparative statics for β_C. First, the derivative of β_C with respect to ϕ_n is more straightforward. From (J.42), we can see that β_C decreases in κ, and from

above we also know that κ decreases in $\tilde{\sigma}_u^2$. Therefore, we have

$$\frac{\partial \beta_C}{\partial \phi_n} = \frac{\partial \beta_C}{\partial \kappa} \frac{\partial \kappa}{\partial \tilde{\sigma}_u^2} \frac{\partial \tilde{\sigma}_u^2}{\partial \phi_n} > 0 \tag{J.48}$$

The derivative of β_C with respect to ϕ_m is more involved. We can compute that

$$\begin{aligned}
\frac{\partial \beta_C}{\partial \phi_m} &= -\frac{1}{\kappa^2} \frac{\partial \kappa}{\partial \phi_m} \left(1 + (1-\lambda) \frac{\lambda (1-\kappa) \rho^2}{1 - \lambda (1-\kappa) \rho^2}\right) + \frac{1-\kappa}{\kappa} \frac{\partial}{\partial \phi_m} \left(1 + (1-\lambda) \frac{\lambda (1-\kappa) \rho^2}{1 - \lambda (1-\kappa) \rho^2}\right) \\
&= -\frac{1}{\kappa^2} \frac{\partial \kappa}{\partial \phi_m} \left(1 + (1-\lambda) \frac{\lambda (1-\kappa) \rho^2}{1 - \lambda (1-\kappa) \rho^2}\right) \\
&\quad + \frac{1-\kappa}{\kappa} (1-\lambda) \frac{\partial}{\partial \phi_m} \left(\frac{\lambda (1-\kappa) \rho^2}{1 - \lambda (1-\kappa) \rho^2}\right) + \frac{1-\kappa}{\kappa} \left(\frac{\lambda (1-\kappa) \rho^2}{1 - \lambda (1-\kappa) \rho^2}\right) \frac{\partial (1-\lambda)}{\partial \phi_m} \\
&= \underbrace{-\frac{1}{\kappa^2} \frac{\partial \kappa}{\partial \phi_m}}_{<0} + \underbrace{\frac{1-\kappa}{\kappa} (1-\lambda) \frac{\partial}{\partial \phi_m} \left(\frac{\lambda (1-\kappa) \rho^2}{1 - \lambda (1-\kappa) \rho^2}\right)}_{<0} \\
&\quad - \frac{1}{\kappa^2} \left(\frac{\lambda (1-\kappa) \rho^2}{1 - \lambda (1-\kappa) \rho^2}\right) \left\{ (1-\lambda) \underbrace{\frac{\partial \kappa}{\partial \phi_m}}_{>0} - k (1-\kappa) \right\}
\end{aligned}$$

The last equation holds because $\frac{\partial (1-\lambda)}{\partial \phi_m} = 1$. Since all the terms except the last one are negatively contributing to $\frac{\partial \beta_C}{\partial \phi_m}$, we can further see that

$$\begin{aligned}
\frac{\partial \beta_C}{\partial \phi_m} &< -\frac{1}{\kappa^2} \frac{\partial \kappa}{\partial \phi_m} + \frac{1-\kappa}{\kappa} \left(\frac{\lambda (1-\kappa) \rho^2}{1 - \lambda (1-\kappa) \rho^2}\right) \\
&= -\frac{1-\kappa}{\kappa} \left(\frac{\partial \sigma^m}{\partial \phi_m} - \frac{\lambda (1-\kappa) \rho^2}{1 - \lambda (1-\kappa) \rho^2}\right) \\
&= -\frac{1-\kappa}{\kappa} \left(\frac{(1-\kappa) \rho^2 \left(\frac{\sigma_z^2}{\sigma} - 1\right)}{1 - \lambda (1-\kappa)^2 \rho^2} - \frac{\lambda (1-\kappa) \rho^2}{1 - \lambda (1-\kappa) \rho^2}\right) \\
&= -\frac{1-\kappa}{\kappa} \frac{\lambda (1-\kappa) \rho^2}{1 - \lambda (1-\kappa) \rho^2} \left\{ \frac{1}{\lambda} \left(\frac{\sigma_z^2}{\Sigma} - 1\right) \frac{1 - \lambda (1-\kappa) \rho^2}{1 - \lambda (1-\kappa)^2 \rho^2} - 1 \right\}
\end{aligned}$$

I would like to show that we can find $\hat{\sigma}_u^2$ such that for any λ, the term in the bracket is positive for all $\tilde{\sigma}_u^2$ such that $\tilde{\sigma}_u^2 \leq \hat{\sigma}_u^2$ and negative otherwise.

First, it is straightforward to see that the term in the bracket is positive for $\tilde{\sigma}_u^2 = 0$ (since

$\Sigma \to 0$) and negative for $\tilde{\sigma}_u^2 \to \infty$ (since $\sigma \to \sigma_z^2$) for any values of ρ, σ_ϵ, and λ. Next, we can also see that the term in the bracket is decreasing in $\tilde{\sigma}_u^2$ for any given ρ, σ_ϵ, and λ: $\frac{\sigma_z^2}{\sigma}$ decreases in $\tilde{\sigma}_u^2$ and $\frac{1-\lambda(1-\kappa)\rho^2}{1-\lambda(1-\kappa)^2\rho^2}$ decreases in $1-k$ (and accordingly also decreases in $\tilde{\sigma}_u^2$). Therefore, there exists a $\hat{\sigma}_u^2$ such that the term in the bracket is positive for any ρ, σ_ϵ, and λ as long as $\sigma_u \leq \hat{\sigma}_u^2$. In practice, we could find such $\hat{\sigma}_u^2$ by finding $\tilde{\sigma}_u^2$ under which

$$\frac{1}{\lambda}\left(\frac{\sigma_z^2}{\Sigma}-1\right)\frac{1-\lambda\left(1-\kappa\right)\rho^2}{1-\lambda\left(1-\kappa\right)^2\rho^2}=1$$

for any given ρ, σ_ϵ^2 and λ. For a given value of ρ and σ_ϵ^2, we can define the minimum $\hat{\sigma}_u^2$ for all λ as $\hat{\sigma}_u^2 \equiv g\left(\rho, \sigma_\epsilon\right)$. Therefore, we can conclude that $\frac{\partial \beta_C}{\partial \phi_m} < 0$ as long as $\tilde{\sigma}_u^2 \leq g\left(\rho, \sigma_\epsilon^2\right)$. Equivalently, $\frac{\partial \beta_C}{\partial \phi_m} < 0$ as long as $\phi_n \leq \bar{\phi}_n \equiv \bar{g}\left(\rho, \sigma_\epsilon^2\right)$, where $\bar{g}\left(\rho, \sigma_\epsilon^2\right)$ can be easily defined using the definition $\tilde{\sigma}_u^2 = \frac{\phi_n}{1-\phi_n}\sigma_z^2$.

10.5 Estimation

Data Source Description

Survey Forecasts Data

The Survey of Professional Forecasters (SPF) began in 1968:Q4 and was taken over by the Philadelphia Fed in 1990:Q2. Forecasters submit their projections in the middle month of each quarter. Two major new data releases are available to the survey participants before submitting their survey. One is the release of the Bureau of Economic Analysis' advance report of the national income and product accounts, which contains the first estimate of GDP and its components for the previous quarter. This is released at the end of the first month of each quarter. The other is the release of the Bureau of Labor Statistics' monthly Employment Situation Report, which is released on the first Friday of each month.

Variable information I use the following eleven variables in the "U.S. Business Indicators" Section. To ease the notation burden, I use the acronym when necessary.

1. Nominal Gross Domestic Product (NGDP)
    - Seasonally adjusted, annual rate
    - Before 1992, forecasts for nominal GNP
2. Real Gross Domestic Product (RGDP)
    - Seasonally adjusted, annual rate
    - Chain-weighted real GDP. Before 1992, fixed-weighted real GDP. Before 1981:Q3, RGDP is computed as NGDP/PGDP*100.
3. GDP Chain-Weighted Price Index (PGDP)
    - Seasonally adjusted, annual rate

- Chain-weighted GDP price index. The base year varies. Before 1992, GNP deflator.

4. Corporate Profits After Taxes (CPROF)

    - Seasonally adjusted, annual rate
    - Before 2006, nominal corporate profits after tax, excluding inventory valuation adjustment (IVA) and capital consumption adjustment (CCAdj)

5. Civilian Unemployment Rate (UNEMP)

    - Seasonally adjusted
    - Quarterly average of the monthly unemployment rates

6. Industrial Production Index (INDPROD)

    - Seasonally adjusted
    - The base year of the index varies
    - Quarterly average of the monthly levels

7. Housing Starts (HOUSING)

    - Seasonally adjusted, annual rate
    - Quarterly average of the monthly levels

8. Consumer Price Index (CPI)

    - Seasonally adjusted
    - Headline CPI inflation rate. The unit of the quarterly forecasts is a quarter-over-quarter annualized growth rate of the quarterly average price index level
    - Survey starts in 1981:Q3

9. 3-month Treasury Bill Rate (TBILL)

- Quarterly average of the daily levels

10. AAA Corporate Bond Yield (BOND)

    - Quarterly average of the daily levels of Moody's Aaa corporate bond yields
    - Before 1990Q4, new, high-grade corporate bond yield

11. 10-year Treasury Bond Rate (TBOND)

    - Quarterly average of the daily levels of 10-year Treasury bond rate

Data availability The survey forecasts have been available for most of these variables since 1968Q4. Exceptions are CPI, TBILL, BOND, and TBOND, whose survey forecasts became available in 1981.

Forecast horizons Forecasters provide (1) quarterly projections for five quarters (current and up to four-quarter-ahead) and (2) annual projections for the current and the following year. For this paper, I use quarterly projections.

Forecast unit Forecasters could provide forecasts using either level or growth rates for most variables. The exception is the forecasts for CPI and PCE, for which forecasters make quarter-over-quarter forecasts.

I compute forecasters' projections about how the variables will change from the previous quarter. For most variables, I take a log difference. For the financial variables and the unemployment, I take the difference. I annualize this difference to compare across different forecast horizons. For example, for the variables I take the log-difference, forecasts are defined as

$$F_{i,t}\, y_{t+h} = \left(\log\left(F_{i,t}\, Y_{t+h}\right) - \log\left(F_{i,t}\, Y_{t-1}\right)\right) \times \frac{4}{h}$$
$$F_{i,t-1}\, y_{t+h} = \left(\log\left(F_{i,t-1}\, Y_{t+h}\right) - \log\left(F_{i,t-1}\, Y_{t-1}\right)\right) \times \frac{4}{h+1}$$

For variables I take the difference, forecasts are defined as

$$F_{i,t}\, y_{t+h} = (F_{i,t}\, Y_{t+h} - F_{i,t}\, Y_{t-1}) \times \frac{4}{h}$$
$$F_{i,t-1}\, y_{t+h} = (F_{i,t-1}\, Y_{t+h} - F_{i,t-1}\, Y_{t-1}) \times \frac{4}{h+1}$$

When computing the forecast revision, I compare these forecasts to those made in the previous quarter. Forecasts from the previous quarters are projections about how the variables will change in the next quarter. Forecast revisions are defined as

$$F_{i,t}\, y_{t+h} - F_{i,t-1}\, y_{t+h}$$

Outlier treatment After constructing the forecasts described above, I drop some observations to restrict the influence of a few outlier variables. First, in each period, I remove forecasts that are five quantiles outsides of the median forecasts. And I only keep individual forecasts that have more than ten observations of the error-revision pairs.

I further restrict samples to measure the forecast behavior in the normal business cycle. During a likely structural change, forecasters might use different forecasting models than the one they would use during the regular cycle. To systematically identify these episodes, I compute the average size of forecast revisions among forecasters each period and remove the top 5 percentile periods. I find such periods of extensive revisions for each forecast horizon. For variables of 200-period observations, I am dropping ten periods. For example, here is the list of periods removed for the forecast of the current quarter realizations for each variable.

1. NGDP: 1974q4, 1975q1, 1980q1, 1981q3, 1981q4, 2001q4, 2008q4, 2009q1, 2020q2, 2020q3

2. RGDP: 1970q4, 1974q4, 1975q1, 1980q1, 1980q2, 1981q4, 2001q4, 2009q1, 2020q2, 2020q3

3. PGDP: 1970q2, 1973q4, 1974q2, 1974q4, 1975q2, 1979q3, 1980q2, 1981q3, 2020q2, 2022q2
4. CPROF: 1974q4, 1981q1, 1981q4, 1982q1, 1982q2, 2002q1, 2005q4, 2020q2, 2020q3 2020q4
5. UNEMP 1974q4, 1975q1, 1980q2, 1981q4, 1982q4, 2001q4, 2009q1, 2009q2, 2020q2, 2020q3, 2020q4
6. INDPROD: 1970q4, 1974q4, 1975q1, 1980q2, 1980q3, 1981q4, 1982q1, 1982q4, 2020q2, 2020q3
7. HOUSING: 1973q4, 1974q4, 1978q2, 1980q2, 1981q1, 1981q3, 1981q4, 2009q1, 2020q2, 2020q3
8. CPI: 1982q1, 1983q1, 1986q2, 1990q4, 2008q4, 2009q1, 2015q1, 2020q2
9. TBILL: 1981q4, 1982q1, 1982q3, 1982q4, 1984q4, 2001q4, 2008q4, 2020q2
10. BOND: 1981q4, 1982q1, 1982q2, 1982q3, 1982q4, 1983q3, 1984q2, 1994q2
11. TBOND: 1992q3, 1994q2, 1996q2, 2002q3, 2008q1, 2020q2, 2022q2

Real-time Macroeconomic Data

I use the real-time data set provided by the Philadelphia Fed. The first release of each variable is used as the "true" realization, which has two uses for my exercise. First, I use this data to compute the forecast errors. Second, I estimate the parameters related to the data-generating process using this data. The last data point I use is 2019:Q4. This is because many variables have an abrupt change during the Covid period, which I assume is not well described as a stationary distribution.

Using real-time data allows us to compute the forecast error correctly. Macroeconomic variables are redefined or reclassified, and the base year changes for the real variables.

Therefore, we must compare the forecast data to a correct realized macro variable with a consistent definition. The real-time data includes the latest data available at any given vintage. The data released for the same vintage is constructed based on an internally consistent variable definition and the same base year. At least for data released after 1996 (when the chain weighting replaced the fixed-weighing method), the change of base year doesn't affect the growth rate of variables.

Regression Estimation

As discussed in the main text, I estimate three regressions. First, following the specification proposed in Bordalo, Gennaioli, Ma, and Shleifer (2020b), I estimate the following regression.

$$y_{t+h} - F_{i,t}\, y_{t+h} = \alpha_{i,I} + \beta_I\ (F_{i,t}\, y_{t+h} - F_{i,t-1}\, y_{t+h}) + I_h + error_{i,t,h} \tag{J.49}$$

$F_{i,t}\, y_{t+h}$ is forecaster i's projected h-quarter-ahead change of y_t from the previous quarter, and the revision variable captures how her belief changed from the previous quarter. $\alpha_{i,I}$ is a dummy variable for each forecaster, and I_h is a dummy variable for each forecast horizon that ranges from $h = 0$ to $h = 3$. I pool all forecast horizons when estimating β_I. The top panel in Table 3.7 reports the results.

The second regression is from Coibion and Gorodnichenko (2015).

$$y_{t+h} - F_t\, y_{t+h} = \alpha_C + \beta_C\ (F_t\, y_{t+h} - F_{t-1}\, y_{t+h}) + I_h + error_{t,h} \tag{J.50}$$

$F_t\, y_{t+h}$ is the average forecast of $F_{i,t}\, y_{t+h}$, for which I use the sample mean of individual forecasts at any given time t. Again, I pool all forecast horizons when estimating β_C. The middle panel in Table 3.7 reports the results.

Finally, the last regression follows the specification from Gemmi and Valchev (2021).

$$F_{i,t}\, y_{t+h} - F_{i,t-1}\, y_{t+h} = \alpha_{i,K} + \beta_K \left(F_{t-1}\, y_{t+h} - F_{i,t-1}\, y_{t+h}\right) + D_t + I_h + error_{i,t} \qquad \text{(J.51)}$$

where D_t is the time dummy. This specification, in essence, regresses the de-meaned forecast revision on de-meaned forecast surprises (defined as $y_{t+h} - F_{i,t-1}\, y_{t+h}$). All forecast horizons are pooled. The bottom panel in Table 3.7 reports the results.

I also report the estimated regression coefficients using only a single forecast horizon. Table 3.8 shows the coefficients estimated from the current quarter forecasts. And Table 3.9 shows the coefficients estimated from the three-quarter-ahead forecasts. Finally, I also report the coefficients using the entire sample period in Table 3.10. For this version, I do not drop the high-mean-squared-error periods identified in the previous section.

CPI long-term forecasts

To estimate the uncertainty about the long-run mean, I estimate how forecasts of μ are revised in response to news about the current quarter. I use the following specification to build on the intuition of Gemmi and Valchev (2021).

$$\left(F_{i,t}\, \mu - F_{i,t-1}\, \mu\right) - \left(F_t\, \mu - F_{t-1}\, \mu\right) = \alpha_{i,\mu,K} + \beta_{\mu,K} \left(F_{t-1}\, \mu - F_{i,t-1}\, \mu\right) + error_{i,\mu,t} \qquad \text{(J.52)}$$

where $F_{i,t}\, \mu$ is the forecast about the long-run, and $F_{i,t}\, \mu$ is the average of $F_{i,t}\, \mu$ across forecasters at time t.

Among the forecast data, the only variable that allows the estimation of the above regression specification is that of the CPI. SPF asks panelists to submit their views about the annual average rate of headline CPI inflation over the next five and ten years. The five-year forecast data started in 2005Q3, and the ten-year forecast data started in 1991Q4. Table 3.11 reports the estimation results. I also report the response of the three-quarter-ahead and the current-quarter forecasts in response to the news about the current quarter's CPI

as a comparison. Unlike the previous regression specifications in Table 3.7, I transform the quarterly forecast data to reflect the annual average inflation rate to maintain the definition consistent with the long-term forecast data.

Table 3.7: Baseline Regression Coefficients

(a) Bordalo et al. Specification

	NGDP	RGDP	PGDP	CPROF	UNEMP	INDPROD	HOUSING	CPI	TBILL	BOND	TBOND
Revision	-0.271***	-0.249***	-0.318***	-0.440***	-0.0476	-0.180***	-0.252***	-0.174**	-0.00757	-0.349***	-0.124**
	(0.0446)	(0.0515)	(0.0445)	(0.0465)	(0.0470)	(0.0607)	(0.0603)	(0.0853)	(0.0390)	(0.0438)	(0.0478)
N	20919	20875	20657	14646	21279	19364	20126	14722	14993	12551	12645

Significance: *=10%, **=5%; ***=1%. Standard errors in parentheses are two-way clustered in forecaster and time.

Dummy variables for forecaster and forecast horizon are controlled. Variables have different sample periods. The longest sample is 1968Q1-2022Q2.

(b) Coibion-Gorodnichenko Specification

	NGDP	RGDP	PGDP	CPROF	UNEMP	INDPROD	HOUSING	CPI	TBILL	BOND	TBOND
Revision	0.632***	0.452***	0.706***	0.685***	0.617***	0.610***	0.505***	0.548***	0.294***	0.0528	0.283***
	(0.106)	(0.129)	(0.129)	(0.181)	(0.0913)	(0.139)	(0.113)	(0.164)	(0.0459)	(0.0719)	(0.0573)
N	797	796	799	791	800	796	795	604	614	613	455

Significance: *=10%, **=5%; ***=1%. Standard errors in parentheses are robust to arbitrary heteroskedasticity and autocorrelation.

Dummy variables for forecast horizon are controlled. Variables have different sample periods. The longest sample is 1968Q1-2022Q2.

(c) Gemmi-Valchev Specification

	NGDP	RGDP	PGDP	CPROF	UNEMP	INDPROD	HOUSING	CPI	TBILL	BOND	TBOND
Surprise	0.545***	0.600***	0.600***	0.513***	0.623***	0.576***	0.579***	0.560***	0.727***	0.678***	0.701***
	(0.0256)	(0.0227)	(0.0277)	(0.0289)	(0.0190)	(0.0239)	(0.0241)	(0.0303)	(0.0319)	(0.0234)	(0.0275)
N	21302	21268	20950	14896	21560	19684	20463	15008	15157	12638	12727

Significance: *=10%, **=5%; ***=1%. Standard errors in parentheses are two-way clustered in forecaster and time.

Dummy variables for time, forecaster and forecast horizon are controlled. Variables have different sample periods. The longest sample is 1968Q1-2022Q2.

Table 3.8: Regression Coefficients for Current-quarter Forecasts Only

(a) Bordalo et al. Specification

	NGDP	RGDP	PGDP	CPROF	UNEMP	INDPROD	HOUSING	CPI	TBILL	BOND	TBOND
Revision	-0.300***	-0.310***	-0.365***	-0.438***	-0.0794*	-0.193***	-0.292***	-0.0795	-0.0742**	-0.327***	-0.0546
	(0.0473)	(0.0536)	(0.0508)	(0.0618)	(0.0448)	(0.0684)	(0.0609)	(0.0762)	(0.0370)	(0.0412)	(0.0394)
N	5346	5357	5289	3722	5305	4918	5155	3757	3794	3183	3175

Significance: *=10%, **=5%; ***=1%. Standard errors in parentheses are two-way clustered in forecaster and time.

Dummy variables for forecaster and forecast horizon are controlled. Variables have different sample periods. The longest sample is 1968Q1-2022Q2.

(b) Coibion-Gorodnichenko Specification

	NGDP	RGDP	PGDP	CPROF	UNEMP	INDPROD	HOUSING	CPI	TBILL	BOND	TBOND
Revision	0.524***	0.243	0.553***	0.879***	0.455***	0.483**	0.369***	0.567**	0.182***	0.0351	0.293***
	(0.159)	(0.185)	(0.199)	(0.264)	(0.0892)	(0.190)	(0.139)	(0.229)	(0.0447)	(0.0921)	(0.0587)
N	202	202	203	201	202	202	202	153	155	155	114

Significance: *=10%, **=5%; ***=1%. Standard errors in parentheses are robust to arbitrary heteroskedasticity and autocorrelation.

Dummy variables for forecast horizon are controlled. Variables have different sample periods. The longest sample is 1968Q1-2022Q2.

(c) Gemmi-Valchev Specification

	NGDP	RGDP	PGDP	CPROF	UNEMP	INDPROD	HOUSING	CPI	TBILL	BOND	TBOND
Surprise	0.591***	0.669***	0.666***	0.533***	0.756***	0.669***	0.673***	0.653***	0.921***	0.851***	0.861***
	(0.0312)	(0.0267)	(0.0321)	(0.0387)	(0.0244)	(0.0298)	(0.0257)	(0.0382)	(0.0405)	(0.0289)	(0.0281)
N	5400	5414	5316	3764	5332	4964	5203	3786	3794	3183	3175

Significance: *=10%, **=5%; ***=1%. Standard errors in parentheses are two-way clustered in forecaster and time.

Dummy variables for time, forecaster and forecast horizon are controlled. Variables have different sample periods. The longest sample is 1968Q1-2022Q2.

Table 3.9: Regression Coefficients for three-quarter-ahead Forecasts Only

(a) Bordalo et al. Specification

	NGDP	RGDP	PGDP	CPROF	UNEMP	INDPROD	HOUSING	CPI	TBILL	BOND	TBOND
Revision	-0.236***	-0.140	-0.206***	-0.461***	0.111	-0.147	-0.285**	-0.302***	0.0998	-0.407***	-0.291***
	(0.0721)	(0.0951)	(0.0629)	(0.0604)	(0.117)	(0.0888)	(0.111)	(0.112)	(0.0873)	(0.0653)	(0.0823)
N	4994	4920	4954	3496	5101	4592	4810	3570	3611	3080	3061

Significance: *=10%, **=5%; ***=1%. Standard errors in parentheses are two-way clustered in forecaster and time.

Dummy variables for forecaster and forecast horizon are controlled. Variables have different sample periods. The longest sample is 1968Q1-2022Q2.

(b) Coibion-Gorodnichenko Specification

	NGDP	RGDP	PGDP	CPROF	UNEMP	INDPROD	HOUSING	CPI	TBILL	BOND	TBOND
Revision	0.977***	0.858**	1.090***	0.181	1.306***	1.002***	0.347	0.453	0.504**	-0.113	0.188
	(0.365)	(0.352)	(0.375)	(0.509)	(0.392)	(0.378)	(0.444)	(0.418)	(0.252)	(0.266)	(0.261)
N	194	194	195	193	195	193	194	149	152	152	113

Significance: *=10%, **=5%; ***=1%. Standard errors in parentheses are robust to arbitrary heteroskedasticity and autocorrelation.

Dummy variables for forecast horizon are controlled. Variables have different sample periods. The longest sample is 1968Q1-2022Q2.

(c) Gemmi-Valchev Specification

	NGDP	RGDP	PGDP	CPROF	UNEMP	INDPROD	HOUSING	CPI	TBILL	BOND	TBOND
Surprise	0.500***	0.509***	0.545***	0.489***	0.439***	0.467***	0.449***	0.466***	0.507***	0.481***	0.491***
	(0.0291)	(0.0341)	(0.0316)	(0.0286)	(0.0195)	(0.0305)	(0.0318)	(0.0332)	(0.0254)	(0.0228)	(0.0282)
N	5131	5059	5060	3579	5212	4706	4931	3683	3691	3130	3115

Significance: *=10%, **=5%; ***=1%. Standard errors in parentheses are two-way clustered in forecaster and time.

Dummy variables for time, forecaster and forecast horizon are controlled. Variables have different sample periods. The longest sample is 1968Q1-2022Q2.

Table 3.10: Regression Coefficients Using All Sample Periods

(a) Bordalo et al. Specification

	NGDP	RGDP	PGDP	CPROF	UNEMP	INDPROD	HOUSING	CPI	TBILL	BOND	TBOND
Revision	-0.114**	-0.130**	-0.230***	-0.393***	-0.278***	-0.149	-0.259***	-0.0499	-0.0226	-0.362***	-0.114**
	(0.0573)	(0.0631)	(0.0554)	(0.0660)	(0.0382)	(0.0931)	(0.0698)	(0.123)	(0.0443)	(0.0446)	(0.0439)
N	21925	21904	21547	15391	22362	20309	21017	15366	15516	12871	13192

Significance: *=10%, **=5%; ***=1%. Standard errors in parentheses are two-way clustered in forecaster and time.

Dummy variables for forecaster and forecast horizon are controlled. Variables have different sample periods. The longest sample is 1968Q1-2022Q2.

(b) Coibion-Gorodnichenko Specification

	NGDP	RGDP	PGDP	CPROF	UNEMP	INDPROD	HOUSING	CPI	TBILL	BOND	TBOND
Revision	0.152*	0.105	0.772***	0.483***	-0.209***	0.318***	0.201	0.463***	0.162***	-0.147**	0.204***
	(0.0835)	(0.0809)	(0.148)	(0.183)	(0.0369)	(0.121)	(0.138)	(0.107)	(0.0536)	(0.0737)	(0.0429)
N	837	837	837	833	841	837	837	638	646	646	478

Significance: *=10%, **=5%; ***=1%. Standard errors in parentheses are robust to arbitrary heteroskedasticity and autocorrelation.

Dummy variables for forecast horizon are controlled. Variables have different sample periods. The longest sample is 1968Q1-2022Q2.

(c) Gemmi-Valchev Specification

	NGDP	RGDP	PGDP	CPROF	UNEMP	INDPROD	HOUSING	CPI	TBILL	BOND	TBOND
Surprise	0.578***	0.631***	0.611***	0.498***	0.783***	0.601***	0.628***	0.589***	0.752***	0.698***	0.704***
	(0.0350)	(0.0277)	(0.0275)	(0.0323)	(0.0782)	(0.0371)	(0.0360)	(0.0408)	(0.0313)	(0.0256)	(0.0267)
N	22308	22297	21916	15641	22643	20629	21354	15652	15680	12976	13354

Significance: *=10%, **=5%; ***=1%. Standard errors in parentheses are two-way clustered in forecaster and time.

Dummy variables for time, forecaster and forecast horizon are controlled. Variables have different sample periods. The longest sample is 1968Q1-2022Q2.

Table 3.11: Estimationg Using the Long-term CPI Forecasts

	10-Year	5-Year	3-quarter	Current quarter
Surprise	0.0828***	0.129***	0.460***	0.855***
	(0.0183)	(0.0413)	(0.0599)	(0.0310)
N	2672	1613	3496	3602

Significance: *=10%, **=5%; ***=1%.
Standard errors in parentheses are two-way clustered in forecaster and time.
Dummy variables for the forecaster and forecast horizon are controlled.
Variables have different sample periods. The longest sample is 1968Q4-2022Q2.

Data Generating Process: AR(1)

I use the following steps to set the parameters for each macroeconomic variable y_t. Using the actual realization, I first get the OLS estimates of the AR(1) parameter ρ and σ_ϵ^2. Table 3.12 reports the parameters.

Table 3.12: Data Generating Process: AR(1) process

	ρ	σ_ϵ
Nominal Gross Domestic Product	0.89	1.31
Real Gross Domestic Product	0.84	1.32
GDP Chain-Weighted Price Index	0.83	1.31
Corporate Profits After Taxes	0.75	8.03
Industrial Production Index	0.85	2.49
Housing Starts	0.84	11.68
Consumer Price Index	0.75	2.0
AAA Corporate Bond Yield	0.83	0.51
Treasury Bond Rate, 10-year	0.82	0.63

10.6 A Stationary Relationship

In summary, the posterior mean for x_t evolves according to

$$\begin{aligned} x_{i,i,t|t} &= (I - K_t)\, x^m_{i,i,t|t} + K_t\, x_t + \bar{\nu}_t + \bar{u}_{i,t} \\ &= (I - K_t)\left((I - \Lambda_t)\,\mu_x + \Lambda_t\, x_{i,i,t|t-1} + \omega_{i,t}\right) + K_t\, x_t + \bar{\nu}_t + \bar{u}_{i,t} \\ &= (I - K_t)\,(I - \Lambda_t)\,\mu_x + (I - K_t)\,\Lambda_t\, A\, x_{i,i,t-1|t-1} + K_t\, x_t + \bar{\nu}_t + \bar{\omega}_{i,t} + \bar{u}_{i,t} \end{aligned}$$

where $\bar{\omega}_{i,t} \equiv (I - K)\,\omega_{i,t}$. To ease the notation burden, I define $\Delta_t \equiv (I - K_t)\,(I - \Lambda_t)$, $\hat{c}_t \equiv \Delta_t\,\mu_x$ and $\hat{A}_t \equiv (I - K_t - \Delta_t)\,A$. Then,

$$x_{i,i,t|t} = \hat{c}_t + \hat{A}_t\, x_{i,i,t-1|t-1} + K_t\, x_t + \bar{\nu}_t + \bar{\omega}_{i,t} + \bar{u}_{i,t} \tag{J.53}$$

Stationary relationship As $t \to \infty$, these matrixes converge to a steady state level. Let's denote this as $\bar{K}_t \to \bar{K}$, $K_t \to K$ and $\Lambda_t \to \Lambda$. Then, forecasts for x_t evolve according to

$$x_{i,i,t|t} = \hat{c} + \hat{A}\, x_{i,i,t-1|t-1} + K\, x_t + \bar{\nu}_t + \bar{\omega}_{i,t} + \bar{u}_{i,t} \tag{J.54}$$

where the variance of the noise is

$$\begin{aligned} V[\bar{\nu}_t] &= V[\bar{K}\,\nu_{i,t}] = \sigma^2_\nu\,\bar{K}\,\bar{K}' \\ V[\bar{u}_{i,t}] &= V[\bar{K}\,u_{i,t}] = \sigma^2_u\,\bar{K}\,\bar{K}' \\ V[\bar{\omega}_{i,t}] &= V[(I - K)\,\omega_{i,t}] = (I - K)\,\sigma_{\omega,t}\,(I - K)' = (I - K)\,(I - \Lambda)\,V[x_{i,i,t|t-1}]\,\Lambda'\,(I - K)' \\ &= \Delta\, V[x_{i,i,t|t-1}]\,(I - K - \Delta)' \end{aligned}$$

Statistical properties of $x_{i,i,t|t}$

Since $x_{i,i,t|t}$ is a conditional expectation of x_t given available information at time t, we can easily see that following holds.

$$Cov\left[x_{i,i,t|t}, x_t\right] = V[x_{i,i,t|t}] = \sigma_x - \Sigma_{t|t}$$
$$Cov\left[x_{i,i,t|t}, x_{i,i,t-1|t-1}\right] = Cov\left[\left(\hat{A} + K\,A\right) x_{i,i,t-1|t-1}, x_{i,i,t-1|t-1}\right] = (I - \Delta)\,A\left(\sigma_x - \Sigma_{t|t}\right)$$

Evolution of the average forecasts for x_t

The average forecasts for x_t evolve according to

$$x_{t|t} = \hat{c} + \hat{A}\,x_{t-1|t-1} + K\,x_t + \bar{\nu}_t \tag{J.55}$$

Therefore, we can see that

$$x_{i,i,t|t} - x_{t|t} = \hat{A}\left(x_{i,i,t-1|t-1} - x_{t-1|t-1}\right) + \bar{\omega}_{i,t} + \bar{u}_{i,t}$$

By iterating backward, we can also see that the difference between $x_{i,i,t|t}$ and $x_{t|t}$ is the history of noise realizations.

$$x_{i,i,t|t} - x_{t|t} = \sum_{j=0}^{\infty} \hat{A}^j\ (\bar{\omega}_{i,t-j} + \bar{u}_{i,t-j}) \equiv NoiseHistory_{i,t}$$

Therefore, the covariance between $x_{t|t}$ and x_t is the same as the covariance between $x_{i,i,t|t}$ and x_t. Also, we can express the variance of $x_{t|t}$ as the variance of $x_{i,i,t|t}$ subtracted by the variance of the history of noises.

$$V[x_{i,i,t|t}] = V[x_{t|t}] + V[NoiseHistory_{i,t}]$$

where $V[NoiseHistory_{i,t}] = V[\sum_{j=0}^{\infty} \hat{A}^j (\bar{\omega}_{i,t} + \bar{u}_{i,t}) \left(\hat{A}^j\right)']$. Finally, the serial correlation of $x_{t|t}$ is derived.

$$\begin{aligned} Cov\left[x_{t|t}, x_{t-1|t-1}\right] &= \hat{A}\, V[x_{t|t}] + K\, A\, Cov\left[x_t, x_{t|t}\right] \\ &= \left(\hat{A} + K\, A\right) V[x_{i,i,t|t}] - \hat{A}\left(I - \hat{A}\right)^{-1} V[\bar{\omega}_{i,t} + \bar{u}_{i,t}] \\ &= Cov\left[x_{i,i,t|t}, x_{i,i,t-1|t-1}\right] - \hat{A}\left(I - \hat{A}\right)^{-1} V[\bar{\omega}_{i,t} + \bar{u}_{i,t}] \end{aligned}$$

Perceived covariance of individual forecast errors and revisions

Given the prior $x_t|m_{i,t} \sim \mathcal{N}\left(x^m_{i,i,t|t},\, \Sigma^m_{t|t}\right)$, posterior distribution of $x_t|s_{i,t}$ is chosen so that it satisfies $Cov\left[x_t - x_{i,i,t|t}, x_{i,i,t|t} - x^m_{i,i,t|t}\right] = O$. However, we are interested in the covariance between the forecast error and revision observed by an econometrician. To see the difference, it is useful to express the law of motion of $x_{i,i,t|t}$ as follows.

$$\begin{aligned} x_{i,i,t|t} &= (I - K)\, x^m_{i,i,t|t} + K\, x_t + \bar{\nu}_t + \bar{u}_{i,t} \\ &= (I - K)\, x_{i,i,t|t-1} + K\, x_t - (I - K)\,(I - \Lambda)\left(x_{i,i,t|t-1} - \mu_x\right) + \bar{\nu}_t + \bar{u}_{i,t} + \bar{\omega}_{i,t} \end{aligned}$$

From this, we can see that

$$\begin{aligned} Cov\left[x_t - x_{i,i,t|t}, x_{i,i,t|t-1}\right] &= Cov\left[E[x_t - x_{i,i,t|t}\big|\, s_{i,t-1}], E[x_{i,i,t|t-1}\big|\, s_{i,t-1}]\right] \\ &= Cov\left[E[x_t - x_{i,i,t|t}\big|\, s_{i,t-1}], x_{i,i,t|t-1}\right] \\ &= (I - K)\,(I - \Lambda)\, V[x_{i,i,t|t-1}] \end{aligned}$$

Therefore,

$$\begin{aligned}
Cov\left[x_t - x_{i,i,t|t}, x_{i,i,t|t} - x_{i,i,t|t-1}\right] &= Cov\left[x_t - x_{i,i,t|t}, \left(x_{i,i,t|t} - x^m_{i,i,t|t}\right) - \left(x_{i,i,t|t-1} - x^m_{i,i,t|t}\right)\right] \\
&= -Cov\left[x_t - x_{i,i,t|t}, x_{i,i,t|t-1} - x^m_{i,i,t|t}\right] \\
&= -Cov\left[x_t - x_{i,i,t|t}, x_{i,i,t|t-1}\right]\ (I-\Lambda)' - Cov\left[x_{i,i,t|t}, \omega_{i,t}\right] \\
&= -\left(I-K\right)\left(I-\Lambda\right)V[x_{i,i,t|t-1}]\left(I-\Lambda\right)' - \left(I-K\right)V[\omega_{i,t}] \\
&= -\left(I-K\right)\left(I-\Lambda\right)V[x_{i,i,t|t-1}] = -\left(I-K\right)\left(\sigma^m_{t|t} - \sigma_{t|t-1}\right)
\end{aligned}$$

the last equality follows from $V[\omega_{i,t}] = \left(I-\Lambda\right)V[x_{i,i,t|t-1}]\Lambda'$. If memory is perfect, we can confirm that forecast error would not be predicted by forecast revision.

Perceived covariance of average forecast errors and revisions

The statistical properties of the average forecast $x_{t|t}$ are determined from the following law of motion.

$$x_{t|t} = \left(I-K\right)x_{t|t-1} + K\,x_t - \left(I-K\right)\left(I-\Lambda\right)\left(x_{t|t-1} - \mu_x\right) + \bar{\nu}_t$$

Rearranging terms yields

$$K\left(x_t - x_{t|t}\right) = \left(I-K\right)\left\{\left(x_{t|t} - x_{t|t-1}\right) + \left(I-\Lambda\right)\left(x_{t|t-1} - \mu_x\right)\right\} - \bar{\nu}_t$$

If K is invertible,

$$\begin{aligned}
&Cov\left[x_t - x_{t|t}, x_{t|t} - x_{t|t-1}\right] \\
&= K^{-1}\left(I-K\right)\left\{V[x_{t|t} - x_{t|t-1}] + \left(I-\Lambda\right)Cov\left[x_{t|t-1}, x_{t|t} - x_{t|t-1}\right]\right\} - K^{-1}V_{noise}
\end{aligned}$$

Correct covariance of forecast errors and revisions

While DM is uncertain about the value of μ, in reality, μ is a fixed parameter. Therefore, the OLS regression of forecast error on revision will asymptotically converge to the covariance of forecast error and revision arising from a fixed parameter. I show here how such statistics differ from the one derived above (where the covariances are averaged across all possible values of μ according to DM's prior about μ). For individual forecasts,

$$\begin{aligned} &Cov\left[x_t - x_{i,i,t|t}, x_{i,i,t|t} - x_{i,i,t|t-1} \middle| \mu\right] \\ &= Cov\left[x_t - x_{i,i,t|t}, x_{i,i,t|t} - x_{i,i,t|t-1}\right] - Cov\left[E[x_t - x_{i,i,t|t} \middle| \mu], E[x_{i,i,t|t} - x_{i,i,t|t-1} \middle| \mu]\right] \end{aligned}$$

Likewise, for average forecasts,

$$\begin{aligned} &Cov\left[x_t - x_{t|t}, x_{t|t} - x_{t|t-1} \middle| \mu\right] \\ &= Cov\left[x_t - x_{t|t}, x_{t|t} - x_{t|t-1}\right] - Cov\left[E[x_t - x_{t|t} \middle| \mu], E[x_{t|t} - x_{t|t-1} \middle| \mu]\right] \end{aligned}$$

Note that the subtracted terms in the above two cases are the same since it must be that $E[x_{i,i,t|t} \big| \mu] = E[x_{t|t} \big| \mu]$ at all t. This term is non-zero because forecasts are biased even in the long run, as DM fails to learn the correct level of μ. Using the fact that forecasts for x_t are stationary, that is $E[x_{i,i,t|t} \big| \mu] = E[x_{i,i,t-1|t-1} \big| \mu]$, we have

$$E[x_{i,i,t|t} \big| \mu] = \left(I - \hat{A}\right)^{-1} (\Delta \, \mu_x + K \, E[x | \mu])$$

Since we can express $E[x_{i,i,t|t} \big| \mu] = cons + D \, E[x_t | \mu]$, where $D \equiv \left(I - \hat{A}\right)^{-1} K$, the correction term can be derived as

$$Cov\left[E[x_t - x_{i,i,t|t} \big| \mu], E[x_{i,i,t|t} - x_{i,i,t|t-1} \big| \mu]\right] = (I - D) \, V[E[x_t | \mu]] \, D' \, (I - A)'$$

Correct variance of revisions

The individual and average forecast revision variances are derived and can be computed using the previously derived stationary relationship.

$$V[x_{i,i,t|t} - x_{i,i,t|t-1}] = V[x_{i,i,t|t}] + V[x_{i,i,t|t-1}] - Cov\left[x_{i,i,t|t}, x_{i,i,t|t-1}\right] - Cov\left[x_{i,i,t|t-1}, x_{i,i,t|t}\right]$$
$$V[x_{t|t} - x_{t|t-1}] = V[x_{t|t}] + V[x_{t|t-1}] - Cov\left[x_{t|t}, x_{t|t-1}\right] - Cov\left[x_{t|t-1}, x_{t|t}\right]$$

The correct variance of the forecast revision (conditional on a fixed μ) is

$$V[x_{i,i,t|t} - x_{i,i,t|t-1}\big|\,\mu] = V[x_{i,i,t|t} - x_{i,i,t|t-1}] - V[E[x_{i,i,t|t} - x_{i,i,t|t-1}\big|\,\mu]]$$
$$V[x_{t|t} - x_{t|t-1}\big|\,\mu] = V[x_{t|t} - x_{t|t-1}] - V[E[x_{t|t} - x_{t|t-1}\big|\,\mu]]$$

The correction term is the same for average and individual forecasts and is derived as follows.

$$V[E[x_{i,i,t|t} - x_{i,i,t|t-1}\big|\,\mu]] = V[(I - A)\,D\,E[x_t|\,\mu]] = (I - A)\,D\,V[E[x_t|\,\mu]]\,D'\,(I - A)'$$

Gemmi-Valchev Proposal

From (J.54) and (J.55), forecast revisions are expressed as follows.

$$x_{i,i,t|t} - x_{i,i,t|t-1} = \hat{c} + K\left(x_t - x_{i,i,t|t-1}\right) - \Delta\, x_{i,i,t|t-1} + \bar{\nu}_t + \bar{\omega}_{i,t} + \bar{u}_{i,t}$$
$$x_{t|t} - x_{t|t-1} = \hat{c} + K\left(x_t - x_{t|t-1}\right) - \Delta\, x_{t|t-1} + \bar{\nu}_t$$

Then, the difference between individual and consensus forecast revisions (in other words, de-meaned individual forecast revisions) can be derived as

$$\left(x_{i,i,t|t} - x_{i,i,t|t-1}\right) - \left(x_{t|t} - x_{t|t-1}\right) = (K + \Delta)\left(x_{t|t-1} - x_{i,i,t|t-1}\right) + \bar{\omega}_{i,t} + \bar{u}_{i,t}$$

The authors propose to estimate the covariance between the de-meaned forecast revisions and the difference between consensus and individual forecasts from the previous period.

$$Cov\left[\left(x_{i,i,t|t}-x_{i,i,t|t-1}\right)-\left(x_{t|t}-x_{t|t-1}\right),x_{t|t-1}-x_{i,i,t|t-1}\right]=(K+\Delta)\,V[x_{t|t-1}-x_{i,i,t|t-1}]$$

Note that this regression coefficient is well-defined only if individual forecasts deviate from the consensus forecasts (that is, when $\phi_n > 0$ or $\phi_m > 0$).

Furthermore, we can see that the perceived covariance (based on DM's prior about μ) is the same as the correct covariance (given a fixed μ), unlike the other covariances I derived earlier.

$$\begin{aligned}&Cov\left[\left(x_{i,i,t|t}-x_{i,i,t|t-1}\right)-\left(x_{t|t}-x_{t|t-1}\right),x_{t|t-1}-x_{i,i,t|t-1}\middle|\,\mu\right]\\&=Cov\left[\left(x_{i,i,t|t}-x_{i,i,t|t-1}\right)-\left(x_{t|t}-x_{t|t-1}\right),x_{t|t-1}-x_{i,i,t|t-1}\right]\end{aligned}$$

This is because the correction term cancels out by de-meaning. That is,

$$\begin{aligned}E[\left(x_{i,i,t|t}-x_{i,i,t|t-1}\right)-\left(x_{t|t}-x_{t|t-1}\right)\big|\,\mu]&=0\\E[x_{t|t-1}-x_{i,i,t|t-1}\big|\,\mu]&=0\end{aligned}$$

This result follows from $E[x_{i,i,t|t}\big|\,\mu]=E[x_{t|t}\big|\,\mu]$ and $E[x_{i,i,t|t-1}\big|\,\mu]=E[x_{t|t-1}\big|\,\mu]$.

To compute the regression coefficient, it remains to derive an expression for $V[x_{t|t-1}-x_{i,i,t|t-1}]$. Note that

$$x_{t|t}-x_{i,i,t|t}=\hat{A}\left(x_{t-1|t-1}-x_{i,i,t-1|t-1}\right)-(\bar{\omega}_{i,t}+\bar{u}_{i,t})$$

Therefore, $V[x_{t|t}-x_{i,i,t|t}]$ satisfies the following fixed-point relation.

$$V[x_{t|t}-x_{i,i,t|t}]=\hat{A}V[x_{t|t}-x_{i,i,t|t}]\hat{A}'+V[\bar{\omega}_{i,t}+\bar{u}_{i,t}]$$

We can then derive

$$V[x_{t|t-1} - x_{i,i,t|t-1}] = A\, V[x_{t|t} - x_{i,i,t|t}]\, A'$$

Finally, for a given c',

$$\beta = \frac{c'\,(K+\Delta)\,V[x_{t|t-1} - x_{i,i,t|t-1}]\,c}{c'\,V[x_{t|t-1} - x_{i,i,t|t-1}]\,c}$$

Furthermore, the variance of the de-meaned forecast revisions is derived as

$$V[\big(x_{i,i,t|t} - x_{i,i,t|t-1}\big) - \big(x_{t|t} - x_{t|t-1}\big)] = (K+\Delta)\,V[x_{t|t-1} - x_{i,i,t|t-1}]\,(K+\Delta)' + V[\bar{\omega}_{i,t} + \bar{u}_{i,t}]$$

and it must be that $V[\,\big(x_{i,i,t|t} - x_{i,i,t|t-1}\big) - \big(x_{t|t} - x_{t|t-1}\big)\big|\,\mu] = V[\big(x_{i,i,t|t} - x_{i,i,t|t-1}\big) - \big(x_{t|t} - x_{t|t-1}\big)]$.

10.7 Monetary Model

I describe a textbook model below, but more details can be found in Gali (2008, Chapter 3).

Household Problem

A representative, infinitely-lived household maximizes the lifetime utility from consumption and labor.

$$E_0 \sum_{t=0}^{\beta} \left[\frac{C_t^{1-\sigma}}{1-\sigma} - \frac{N_t^{1+\varphi}}{1+\varphi} \right]$$

where C_t is the quantity of the basket of goods consumed at time t, and N_t is the number of hours worked. The consumption/savings and labor-supply decisions are subject to the budget constraint that should be met every period.

$$P_t\, C_t + Q_t\, B_t \leq B_{t-1} + W_t\, N_t + T_t$$

where P_t is the aggregate price index, B_t is the one-period bond and Q_t its price, W_t is the nominal hourly wage, and finally T_t is a lump-sum income. The household should also be solvent after all, which is captured by the condition that $\lim_{T\to\infty} Et\, B_t \geq 0$.

The first order conditions and their Taylor expansion around the zero-inflation steady state imply

$$w_t - p_t = \sigma\, c_t + \varphi\, n_t \tag{J.56}$$

$$c_t = E_t\, c_{t+1} - \frac{1}{\sigma} \left(-q_t - E_t\, \pi_{t+1} + log\, \beta \right) \tag{J.57}$$

where the lowercase denotes the log of the variable denoted in uppercase.

Firm Problem

A continuum of firms indexed by $i \in [0,1]$ produces a differentiated goods. The production function is described as

$$Y_t(i) = A_t \, N_t(i)$$

where A_t is the level of production technology, assumed to be common to all firms and evolve exogenously over time.

Each firm reconsiders its price with probability $1-\alpha$, independent of when its price is readjusted in the past. Thus, at any period, a mass of $1-\alpha$ firms resets their prices and the remaining mass of α firms keep their old prices. The aggregate price index is then formed according to

$$P_t = \left[\alpha \left(P_{t-1} \right)^{1-\eta} + (1-\eta) \left(\int P_{i,t}^{*} \, di \right)^{1-\eta} \right]^{\frac{1}{1-\eta}}$$

Optimal Price Setting

Suppose firm i chooses the price $P_{i,t}^{*}$ in period t. This price maximizes the current market value of the profits if the firm cannot reoptimize the price forever.

$$\max_{P_{i,t}} \quad E_{i,t} \left[\sum_{h=0}^{\infty} \alpha^h \, Q_{t,t+h} \left(P_{i,t} \, Y_{i,t+h|t} - \Psi_{t+h} \left(Y_{i,t+h|t} \right) \right) \right]$$

where α the probability of not resetting prices, $Q_{t,t+h}$ is the stochastic discount factor for evaluating the future nominal payoffs generated at $t+h$, $Y_{i,t+h|t}$ is the output demanded in period $t+h$ if the price remains the one chosen at time t, and Ψ_{t+h} is the (nominal) cost

function at time $t+h$. Firm i takes into account that the demand $Y_{i,t+h|t}$ is given as

$$Y_{i,t+h|t} = \left(\frac{P_{i,t}}{P_{t+h}}\right)^{\eta} C_{t+h}$$

where θ is the elasticity of substitution among goods, P_{t+k} is the aggregate price at time $t+h$ and C_{t+h} is the aggregate consumption at time $t+h$.

The first-order condition implies that

$$E_{i,t}\left[\sum_{h=0}^{\infty} \alpha^h Q_{t,t+h} Y_{i,t+h|t} \left(P^*_{i,t} - \mathcal{M}\psi_{t+h}\right)\right] = 0$$

where $\mathcal{M} \equiv \frac{\eta}{\eta-1}$ and ψ_{t+h} is the nominal marginal cost at $t+h$. Dividing by P_{t-1} and letting $\Pi_{t,t+h} \equiv \frac{P_{t+h}}{P_t}$, we can rewrite the first order condition as

$$E_{i,t}\left[\sum_{h=0}^{\infty} \alpha^h Q_{t,t+h} Y_{i,t+h|t} \left(\frac{P^*_{i,t}}{P_{t-1}} - \mathcal{M} MC_{t+h}\, \Pi_{t,t+h}\right)\right] = 0$$

First-order Taylor expansion around the zero-inflation steady state implies that

$$\begin{aligned} p^*_{i,t} - p_{t-1} &= E_{i,t}\left[(1-\alpha\beta)\sum_{h=0}^{\infty} (\alpha\beta)^h \left((mc_{t+h} - mc) + (p_{t+h} - p_{t-1})\right)\right] \\ &= E_{i,t}\left[\sum_{h=0}^{\infty} (\alpha\beta)^h \left\{(1-\alpha\beta)(mc_{t+h} - mc) + \pi_{t+h}\right\}\right] \end{aligned}$$

where mc is the steady state value of mc_{t+h}. From this expression, we can see that the optimal reset price $p^*_{i,t}$ equals mc over a weighted average of the current and expected nominal marginal costs.

Note that the marginal cost at $t+h$ does not depend on the quantity firm i supplies. This is because the marginal product of labor does not depend on quantity, as $mpn_t = a_t$.

Thus,

$$mc_{t+h} = w_{t+h} - p_{t+h} - mpn_{t+h} = w_{t+h} - p_{t+h} - a_{t+h}$$

Equilibrium

Since market clears for all i goods, it follows that

$$C_t = Y_t$$

which implies $c_t = y_t$. And the labor market clears, requiring

$$N_t = \int N_t(i)\, di$$

which can be shown to imply $n_t = y_t - a_t$ in the first order approximation. Thus, using the household's optimality condition,

$$w_t - p_t = (\sigma + \varphi)\, y_t - \varphi\, a_t$$

Denoting y_t^n as the efficient level of output, we can show that $y_t^n = \frac{1+\varphi}{\sigma+\varphi}\, a_t$. I define the output gap as

$$x_t = y_t - y_t^n$$

Thus, the marginal costs are derived as

$$mc_{t+h} = (\sigma + \varphi)\, x_t$$

Firms' Macroeconomic Expectations

Substituting (G.22), we can see that inflation is determined as

$$\pi_t = (1-\alpha)\left(\hat{z}_t + \frac{\alpha\beta}{1-\alpha\beta}\,\hat{\mu}_t\right)$$

Substituting (G.23) and (G.24), we get

$$\pi_t = (1-\alpha)\left\{(\kappa+\kappa_\mu)\,z_t + \left(\lambda(1-\kappa) + \frac{\alpha\beta}{1-\alpha\beta}\lambda(1-\kappa_\mu)\right)\hat{\mu}_{t-1}\right\}$$

Defining $\hat{\kappa} = \kappa + \kappa_\mu$ and $\hat{b} = \lambda(1-\kappa) + \frac{\alpha\beta}{1-\alpha\beta}\lambda(1-\kappa_\mu)$, we can describe the above expression as

$$\pi_t = (1-\alpha)\left\{\hat{\kappa}\,z_t + \hat{b}\,\hat{\mu}_{t-1}\right\} \tag{J.58}$$

Inflation Determination

We can solve for the equilibrium inflation process using a guess-and-verify approach. The equation (G.21) states that z_t is determined by π_t and e_t, and the equation (J.58) states that π_t is determined by z_t and $\hat{\mu}_{t-1}$. Thus, it is straightforward to see that two state variables, e_t and $\hat{\mu}_{t-1}$, determine inflation, and the relationship is linear. We guess the following inflation process.

$$\pi_t = \varphi_e\, e_t + \varphi_\mu\, \hat{\mu}_{t-1} \tag{J.59}$$

Combining (G.21), (J.58), and (J.59), we can find the coefficients φ_e and φ_μ that verify our initial guess. They are derived as below.

$$\varphi_e = \frac{\delta}{1 + \delta\,\sigma\,s + \frac{1}{\alpha}\,\frac{1-\hat{\kappa}}{\hat{\kappa}}}$$
$$\varphi_\mu = \frac{1}{1 + \delta\,\sigma\,s + \frac{1}{\alpha}\,\frac{1-\hat{\kappa}}{\hat{\kappa}}}\,\frac{1-\alpha}{\alpha}\,\frac{\hat{b}}{\hat{\kappa}}$$

where $\delta \equiv \frac{(1-\alpha)(1-\alpha\beta)}{\alpha}$, $\hat{\kappa} = \kappa + \kappa_\mu$, and $\hat{b} = \lambda(1-\kappa) + \frac{\alpha\beta}{1-\alpha\beta}\lambda(1-\kappa_\mu)$. When exploring alternative monetary policies, I consider values of s such as

$$s = s^* \cdot \frac{\theta}{1-\theta}$$

in which case, we could express the coefficients φ_e and φ_μ as

$$\varphi_e = \frac{1-\theta}{1-\theta+\theta\,\hat{\delta}\,s^* + \frac{1-\theta}{\alpha}\frac{1-\hat{\kappa}}{\hat{\kappa}}}\,\delta$$
$$\varphi_\mu = \frac{1-\theta}{1-\theta+\theta\,\hat{\delta}\,s^* + \frac{1-\theta}{\alpha}\frac{1-\hat{\kappa}}{\hat{\kappa}}}\,\frac{1-\alpha}{\alpha}\,\frac{\hat{b}}{\hat{\kappa}}$$

This expression makes it clear that a complete inflation stabilization ($\theta = 1$) is supported by $\varphi_e = \varphi_m = 0$.

Variability of Inflation

From (J.59), we can see that the variability of inflation is derived as

$$V[\pi_t] = \varphi_e^2\,V[e_t] + \varphi_\mu^2\,V[\hat{\mu}_{t-1}]$$

Therefore, it remains to derive the variability of $\hat{\mu}_t$. First, note that from (G.21), z_t is also determined by two state variables.

$$z_t = \varpi_e \, e_t + \varpi_\mu \, \hat{\mu}_{t-1}$$

where ϖ_e and ϖ_μ are defined as

$$\begin{aligned}
\varpi_e &= \frac{\alpha}{1-\alpha}\left(\left(\frac{1}{\alpha} - \frac{1-\theta+\theta\hat{\delta}}{1-\theta}\right)\varphi_e + \delta\right) \\
\varpi_\mu &= \frac{\alpha}{1-\alpha}\left(\frac{1}{\alpha} - \frac{1-\theta+\theta\hat{\delta}}{1-\theta}\right)\varphi_\mu
\end{aligned}$$

Using this expression, we can then describe the law of motion of $\hat{\mu}_t$ as

$$\hat{\mu}_t = \underbrace{\left(\lambda\left(1-\kappa_\mu\right) + \kappa_\mu \, \varpi_\mu\right)}_{\equiv \rho_\mu} \hat{\mu}_{t-1} + \kappa_\mu \, \varpi_e \, e_t$$

From this, we can see that

$$V[\hat{\mu}_t] = \frac{\left(\kappa_\mu \, \varpi_e\right)^2}{1-\rho_\mu^2} \, V[e_t]$$

Therefore, the variability of inflation is derived as

$$V[\pi_t] = \left(\varphi_e^2 + \frac{\left(\kappa_\mu \, \varpi_e\right)^2}{1-\rho_\mu^2}\right) V[e_t]$$

Bibliography

ADAM, K., A. MARCET, AND J. BEUTEL (2017): "Stock Price Booms and Expected Capital Gains," *American Economic Review*, 107, 2352–2408.

AFROUZI, H., S. Y. KWON, A. LANDIER, Y. MA, AND D. THESMAR (2020): "Overreaction in Expectations: Evidence and Theory," SSRN Working Paper.

AFROUZI, H. AND C. YANG (2021): "Dynamic Rational Inattention and the Phillips Curve," Working Paper.

ANDRADE, P., R. K. CRUMP, S. EUSEPI, AND E. MOENCH (2016): "Fundamental disagreement," *Journal of Monetary Economics*, 83, 106–128.

ANDRADE, P. AND H. LE BIHAN (2013): "Inattentive professional forecasters," *Journal of Monetary Economics*, 60, 967–982.

ANGELETOS, G.-M., F. COLLARD, AND H. DELLAS (2020): "Business-Cycle Anatomy," *American Economic Review*, 110, 3030–3070.

ANGELETOS, G.-M. AND Z. HUO (2021): "Myopia and Anchoring," *American Economic Review*, 111, 1166–1200.

ANGELETOS, G.-M., Z. HUO, AND K. A. SASTRY (2021): "Imperfect Macroeconomic Expectations: Evidence and Theory," *NBER Macroeconomics Annual*, 35, 1–86.

ANGELETOS, G.-M. AND J. LA'O (2013): "Sentiments," *Econometrica*, 81, 739–779.

ANGELETOS, G.-M. AND C. LIAN (2018): "Forward Guidance without Common Knowledge," *The American Economic Review*, 108, 2477–2512.

AZEREDO DA SILVEIRA, R., Y. SUNG, AND M. WOODFORD (2020): "Optimally Imprecise Memory and Biased Forecasts," Working Paper 28075, National Bureau of Economic Research.

BAKER, S., T. MCELROY, AND X. SHENG (2020): "Expectation Formation Following Large, Unexpected Shocks," *The Review of Economics and Statistics*, 102, 287–303.

BAKSHI, G. AND G. SKOULAKIS (2010): "Do subjective expectations explain asset pricing puzzles?" *Journal of Financial Economics*, 98, 462–477.

BARBERIS, N., R. GREENWOOD, L. JIN, AND A. SHLEIFER (2015): "X-CAPM: An extrapolative capital asset pricing model," *Journal of Financial Economics*, 115, 1–24.

BASSETTI, F., R. CASARIN, AND M. DEL NEGRO (2023): "Chapter 15 - Inference on probabilistic surveys in macroeconomics with an application to the evolution of uncertainty in the survey of professional forecasters during the COVID pandemic," in *Handbook of Economic Expectations*, ed. by R. Bachmann, G. Topa, and W. van der Klaauw, Academic Press, 443–476.

BEECHEY, M. J. AND J. H. WRIGHT (2009): "The high-frequency impact of news on long-term yields and forward rates: Is it real?" *Journal of Monetary Economics*, 56, 535–544.

BEN-DAVID, I., E. FERMAND, C. M. KUHNEN, AND G. LI (2018): "Expectations Uncertainty and Household Economic Behavior," Working Paper 25336, National Bureau of Economic Research.

BERARDI, M. AND J. GALIMBERTI (2017): "Empirical Calibration of Adaptive Learning," *Journal of Economic Behavior and Organization*.

BIANCHI, F., S. C. LUDVIGSON, AND S. MA (2022): "Belief Distortions and Macroeconomic Fluctuations," Working Paper.

BOLLERSLEV, T., G. TAUCHEN, AND H. ZHOU (2009): "Expected Stock Returns and Variance Risk Premia," *The Review of Financial Studies*, 22, 4463–4492.

BORDALO, P., J. J. CONLON, N. GENNAIOLI, S. Y. KWON, AND A. SHLEIFER (2022a): "Memory and Probability," *The Quarterly Journal of Economics*.

BORDALO, P., N. GENNAIOLI, R. LA PORTA, AND A. SHLEIFER (2020a): "Expectations of Fundamentals and Stock Market Puzzles," NBER Working Paper No. 27283.

BORDALO, P., N. GENNAIOLI, R. LAPORTA, AND A. SHLEIFER (2022b): "Belief Over-Reaction and Stock Market Puzzles," Working Paper.

BORDALO, P., N. GENNAIOLI, Y. MA, AND A. SHLEIFER (2020b): "Overreaction in Macroeconomic Expectations," *American Economic Review*, 110, 2748–2782.

BORDALO, P., N. GENNAIOLI, R. L. PORTA, AND A. SHLEIFER (2019): "Diagnostic Expectations and Stock Returns," *The Journal of Finance*, 74, 2839–2874.

BOUCHAUD, J.-P., P. KRÜGER, A. LANDIER, AND D. THESMAR (2019): "Sticky Expectations and the Profitability Anomaly," *The Journal of Finance*, 74, 639–674.

BRANCH, W. A. AND G. W. EVANS (2006): "A Simple Recursive Forecasting Model," *Economics Letters*.

BROER, T. AND A. N. KOHLHAS (2021): "Forecaster (Mis-)Behavior," Working Paper.

CAPLIN, A., M. DEAN, AND J. LEAHY (2022): "Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy," *Journal of Political Economy*.

CARVALHO, C., S. EUSEPI, E. MOENCH, AND B. PRESTON (2022): "Anchored Inflation Expectations," Working Paper.

CASE, K. E., R. J. SHILLER, AND A. THOMPSON (2012): "What Have They Been Thinking? Home Buyer Behavior in Hot and Cold Markets," NBER Working Paper No. 18400.

CIESLAK, A. (2018): "Short-Rate Expectations and Unexpected Returns in Treasury Bonds," *The Review of Financial Studies*, 31, 3265–3306.

COIBION, O., U. AUSTIN, Y. GORODNICHENKO, AND M. WEBER (2021a): "MONETARY POLICY COMMUNICATIONS AND THEIR EFFECTS ON HOUSEHOLD INFLATION EXPECTATIONS," NBER Working Paper No. 25482.

COIBION, O., D. GEORGARAKOS, Y. GORODNICHENKO, G. KENNY, AND M. WEBER (2021b): "The Effect of Macroeconomic Uncertainty on Household Spending," Working Paper 28625, National Bureau of Economic Research.

COIBION, O. AND Y. GORODNICHENKO (2012): "What Can Survey Forecasts Tell Us about Information Rigidities?" *Journal of Political Economy*, 120, 116–159.

——— (2015): "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts," *The American Economic Review*, 105, 2644–2678.

COIBION, O., Y. GORODNICHENKO, AND S. KUMAR (2018): "How Do Firms Form Their Expectations? New Survey Evidence," *American Economic Review*, 108, 2671–2713.

COLLIN-DUFRESNE, P., M. JOHANNES, AND L. A. LOCHSTOER (2016): "Parameter Learning in General Equilibrium: The Asset Pricing Implications," *American Economic Review*, 106, 664–698.

——— (2017): "Asset Pricing When 'This Time Is Different'," *Review of Financial Studies*, 30, 505–535.

COVER, T. M. AND J. A. THOMAS (2006): "Elements of Information Theory," *New York: Wiley, 2d ed.*

D'ARIENZO, D. (2020): "Maturity Increasing Overreaction and Bond Market Puzzles," SSRN Working Paper.

DEW-BECKER, I., S. GIGLIO, A. LE, AND M. RODRIGUEZ (2017): "The price of variance risk," *Journal of Financial Economics*, 123, 225–250.

EHLING, P., A. GRANIERO, AND C. HEYERDAHL-LARSEN (2018): "Asset Prices and Portfolio Choice with Learning from Experience," *Review of Economic Studies*.

ELLIOTT, G., I. KOMUNJER, AND A. TIMMERMANN (2008): "Biases in Macroeconomic Forecasts: Irrationality or Asymmetric Loss?" *Journal of the European Economic Association*, 6, 122–157.

ENKE, B. AND T. GRAEBER (2019): "Cognitive Uncertainty," NBER Working Paper No. 26518.

EUSEPI, S. AND B. PRESTON (2011): "Expectations, Learning, and Business Cycle Fluctuations," *American Economic Review*, 101, 2844–2872.

EVANS, G. AND S. HONKAPOHJA (2011): "Learning as a Rational Foundation for Macroeconomics and Finance," CEPR Discussion Papers 8340, C.E.P.R. Discussion Papers.

EVANS, G. W. AND S. HONKAPOHJA (2001): "Learning and Expectations in Macroeconomics," *Princeton University Press*.

FARMER, L., E. NAKAMURA, AND J. STEINSSON (2021): "Learning About the Long Run," .

FUSTER, A., B. HÉBERT, AND D. LAIBSON (2011): "Natural Expectations, Macroeconomic Dynamics, and Asset Pricing," *NBER Macroeconomics Annual*.

FUSTER, A., D. LAIBSON, AND B. MENDEL (2010): "Natural Expectations and Macroeconomic Fluctuations," *The Journal of Economic Perspectives*, 24, 67–84.

GABAIX, X., P. GOPIKRISHNAN, V. PLEROU, AND H. E. STANLEY (2006): "Institutional Investors and Stock-Market Volatility," *Quarterly Journal of Economics*.

GALÍ, J. (2008): *Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework*, Princeton University Press.

GÀTI, L. (2021): "Monetary Policy & Anchored Expectations An Endogenous Gain Learning Model," Working Paper.

GEMMI, L. AND R. VALCHEV (2021): "Biased Surveys," Working Paper.

GIGLIO, S. AND B. KELLY (2018): "Excess Volatility: Beyond Discount Rates," *The Quarterly Journal of Economics*, 133, 71–127.

GREENWOOD, R. AND A. SHLEIFER (2014): "Expectations of Returns and Expected Returns," *Review of Financial Studies*, 27, 714–746.

HAZELL, J., J. HERREÑO, E. NAKAMURA, AND J. STEINSSON (2022): "The Slope of the Phillips Curve: Evidence from U.S. States," *The Quarterly Journal of Economics*, 137.

HOGARTH, R. AND H. EINHORN (1992): "Order Effects in Belief Updating: The Belief-Adjustment Model," *Cognitive Psychology*.

JUODIS, A. AND S. KUCINSKAS (2019): "Quantifying Noise," SSRN Working Paper.

K. CRUMP, R., S. EUSEPI, E. MOENCH, AND B. PRESTON (2021): "The Term Structure of Expectations," Federal Reserve Bank of New York Staff Reports.

KOHLHAS, A. N. AND A. WALTHER (2021): "Asymmetric Attention," *American Economic Review*, 111, 2879–2925.

KOZLOWSKI, J., L. VELDKAMP, AND V. VENKATESWARAN (2020): "The Tail That Wags the Economy: Beliefs and Persistent Stagnation," *Journal of Political Economy*, 128, 2839–2879.

LOCHSTOER, L. A. AND T. MUIR (2022): "Volatility Expectations and Returns," *The Journal of Finance*, 77, 1055–1096.

MACKOWIAK, B. AND M. WIEDERHOLT (2009): "Optimal Sticky Prices under Rational Inattention," *American Economic Review*, 99, 769–803.

MALMENDIER, U. AND S. NAGEL (2016): "Learning from Inflation Experiences," *Quarterly Journal of Economics*.

MALMENDIER, U., D. POUZO, AND V. VANASCO (2020): "Investor Experiences and Financial Market Dynamics," *Journal of Financial Economics*.

MANKIW, N. G. AND R. REIS (2002): "Sticky Information versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve," *The Quarterly Journal of Economics*, 117, 1295–1328.

MANSKI, C. F. (2018): "Survey Measurement of Probabilistic Macroeconomic Expectations: Progress and Promise," *NBER Macroeconomics Annual*, 32, 411–471.

MCELROY, T. AND X. S. SHENG (2021): "Augmented Information Rigidity Test," SSRN Working Paper.

MIAO, J., J. WU, AND E. R. YOUNG (2022): "Multivariate Rational Inattention," *Econometrica*, 90, 907–945.

MILANI, F. (2007): "Expectations, Learning and Macroeconomic Persistence," *Journal of onetary Economics*.

——— (2014): "Learning and Time-varying Macroeconomic Volatility," *Journal of Economic Dynamics and Control*.

MOLAVI, P. (2022): "Simple Models and Biased Forecasts," Working Paper.

NAGEL, S. AND Z. XU (2021): "Asset Pricing with Fading Memory," *The Review of Financial Studies*, 00, 56.

NAKOV, A. AND G. NUÑO (2015): "Learning from Experience in the Stock Market," *Journal of Economic Dynamics and Control*.

NELIGH, N. (2022): "Rational Memory with Decay," Working Paper.

NORDHAUS, W. D. (1987): "Forecasting Efficiency: Concepts and Applications," *The Review of Economics and Statistics*, 69, 667–674.

ORPHANIDES, A. AND J. WILLIAMS (2004): *Imperfect Knowledge, Inflation Expectations, and Monetary Policy*, University of Chicago Press.

ORPHANIDES, A. AND J. C. WILLIAMS (2006): "Monetary Policy with Imperfect Knowledge," *Journal of the European Economic Association*, 4, 366–375.

PATTON, A. J. AND A. TIMMERMANN (2010): "Why do forecasters disagree? Lessons from the term structure of cross-sectional dispersion," *Journal of Monetary Economics*, 57, 803–820.

PESARAN, M. H. AND M. WEALE (2006): "Chapter 14 Survey Expectations," in *Handbook of Economic Forecasting*, Elsevier, vol. 1, 715–776.

PIAZZESI, M. AND M. SCHNEIDER (2009): "Trend and Cycle in Bond Premia," Working Paper.

PRAT-CARRABIN, A., F. MEYNIEL, AND R. A. D. SILVEIRA (2022): "Resource-Rational Account of Sequential Effects in Human Prediction," *bioRXiv*.

PRAT-CARRABIN, A., F. MEYNIEL, M. TSODYKS, AND R. A. D. SILVEIRA (2021): "Biases and Variability from Costly Bayesian Inference," *Entropy*.

REIS, R. (2020): "The People versus the Markets: A Parsimonious Model of Inflation Expectations," CEPR Discussion Papers No. 15624.

ROTEMBERG, J. AND M. WOODFORD (1997): *An Optimization-Based Econometric Framework for the Evaluation of Monetary Policy*, National Bureau of Economic Research, Inc, 297–361.

SCHRAEDER, S. (2016): "Information Processing and Non-Bayesian Learning in Financial Markets," *Review of Finance*.

SIMS, C. A. (2003): "Implications of rational inattention," *Journal of Monetary Economics*, 50, 665–690.

SINGLETON, K. J. (2021): "Presidential Address: How Much "Rationality" Is There in Bond-Market Risk Premiums?" *The Journal of Finance*, 76, 1611–1654.

SLOBODYAN, S. AND R. WOUTERS (2012): "Learning in an Estimated Medium-scale DSGE Model," *Journal of Economic Dynamics and Control*.

STELLA, A. AND J. H. STOCK (2015): "14A state-dependent model for inflation forecasting," in *Unobserved Components and Time Series Econometrics*, Oxford University Press.

SUNG, Y. (2022): "Macroeconomic Expectations and Cognitive Noise," Working Paper.

TIMMERMANN, A. G. (1993): "How Learning in Financial Markets Generates Excess Volatility and Predictability in Stock Prices," *The Quarterly Journal of Economics*, 108, 1135–1145.

WACHTER, J. A. AND M. J. KAHANA (2021): "A retrieved-context theory of financial decisions," SSRN Working Paper.

WACHTER, J. A. AND U. MALMENDIER (2022): "Memory of Past Experiences and Economic Decisions," Working Paper.

WANG, C. (2021): "Under- and Overreaction in Yield Curve Expectations," SSRN Working Paper.

WEITZMAN, M. L. (2007): "Subjective Expectations and Asset-Return Puzzles," *American Economic Review*, 97, 1102–1130.

WOODFORD, M. (2003): "Imperfect common knowledge and the effects of monetary policy," *Knowledge, information, and expectations in modern macroeconomics : in honor of Edmund S. Phelps*, 25–58, Princeton University Press.

——— (2020): "Modeling Imprecision in Perception, Valuation, and Choice," *Annual Review of Economics*, 12, 579–601.

YU, A. J. AND J. D. COHEN (2009): "Sequential Effects: Superstition or Rational Behavior?" *Advances in Neural Information Processing Systems*.

ZARNOWITZ, V. (1985): "Rational Expectations and Macroeconomic Forecasts," *Journal of Business & Economic Statistics*, 3, 293–311.