For Online Publication

Supplemental Appendix

Rava Azeredo da Silveira, Yeji Sung, and Michael Woodford,
"Optimally Imprecise Memory and Biased Forecasts"

A. Reduction of the General Forecasting Problem to Estimation of μ

In the main text, we suppose that DM's forecasts and memory structure minimize the expected loss function (I.2). In this section, we show that the optimization problem can be restricted to a problem of estimating μ, in which the memory system minimizes the discounted sum of mean squared errors in the estimation task.

Consider the problem of choosing the vector of forecasts z_t each period so as to minimize (I.2). 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 (I.2) involving z_t can be equivalently expressed as[53]

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

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 (I.2) can thus equivalently be written as the discounted sum of the L_{1t} terms, which involve squared differences between z_t and $\mathrm{E}_t\tilde{z}_t$.

[53] Here we omit the factor β^t that multiplies this term in (I.2).

It further follows from the law of motion (I.1) 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 (I.2) 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

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

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

$$\text{(A.1)} \qquad \tilde{L}_1(s_t) \;=\; (\hat{z}_t - a\mathrm{E}[\mu|s_t])' W (\hat{z}_t - a\mathrm{E}[\mu|s_t]) \,+\, a'Wa \cdot \mathrm{var}[\mu|s_t],$$

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 (A.1), 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 (I.2) will equal (I.4), 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 (I.2) must be the one that minimizes (I.4); and since any other kind of forecasting rule can only lead to a higher value of (I.2), we can replace the problem of choosing a rule for determining z_t that minimizes (I.2) by the problem of choosing a rule for determining $\hat{\mu}_t$ that minimizes (I.4).

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

In this section, we show how DM's belief is updated given the feasible class of memory system assumed in (I.5). We discuss the Kalman Filter problem when the external state y_t is observed.As in section III, we define the state vector as $x_t \equiv (\mu,\, y_{t-1})$. Given any inherited memory state m_t, we partition its elements as as

$$m_t \;=\; \begin{bmatrix} \underline{m}_t \\ \bar{m}_t \end{bmatrix}, \tag{B.1}$$

where the lower block consists of the elements of the "reduced" memory state, $\bar{m}_t \equiv \mathrm{E}[x_t\,|m_t]$, 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.)

The assumed memory structure implies that a posterior distribution of x_t conditional on the memory state m_t is of the form

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

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

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

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

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

where the vector $c' \equiv [1 - \rho\,\rho]$ is from (I.1) and 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,

$$x_t \,|\bar{m}_t, y_t \;\sim\; N(\bar{\mu}_t,\, \bar{\Sigma}_t), \tag{B.3}$$

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 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

$$\gamma_t \;=\; \frac{\Sigma_t c}{c'\Sigma_t c + \sigma_\epsilon^2}. \tag{B.4}$$

The gain coefficient γ_{1t} in equation (II.2) is just the first element of this vector, $\gamma_{1t} \equiv e_1'\gamma_t$.

The variance-covariance matrix in the conditional distribution (B.3) 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

$$\hat{\sigma}_t^2 \;=\; e_1'\bar{\Sigma}_t e_1 \;=\; e_1'\Sigma_t e_1 \;-\; (c'\Sigma_t c + \sigma_\epsilon^2)(\gamma_{1t})^2. \tag{B.5}$$

Substituting expression (B.2) 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 (I.8). It remains to be shown that this is an upper bound for $\hat{\sigma}_t^2$. To show this, we observe that

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

This establishes the upper bound (I.8) stated in the main text.

C. Information Optimally Recorded in the Memory Structure

In this section, we derive that the optimal memory structure records information only about the "reduced" cognitive state, as represented in (III.4). In our general analysis, the reduced cognitive state is defined as

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

Note that in the simple case discussed in section II, the recorded reduced cognitive state is simply $\bar{s}_t = \hat{\mu}_t$. This simplification arises because y_t is a transitory process: it is only the knowledge about $\hat{\mu}_t$ that increases accuracy of forecasts that will be made in time $t+1$ and beyond. In this case, we wish to show that the optimal memory structure records information only about $\hat{\mu}_t$, as represented in (II.4). The derivation below applies analogously.

Let the feasible memory structure (I.5) be written in the partitioned form

$$\text{(C.1)} \qquad \begin{bmatrix} \underline{m}_{t+1} \\ \bar{m}_{t+1} \end{bmatrix} \;=\; \begin{bmatrix} \Lambda_{a,t} & \Lambda_{b,t} \\ \Lambda_{c,t} & \Lambda_{d,t} \end{bmatrix} \begin{bmatrix} \underline{s}_t \\ \bar{s}_t \end{bmatrix} + \begin{bmatrix} \underline{\omega}_{t+1} \\ \bar{\omega}_{t+1} \end{bmatrix}.$$

Here m_{t+1} is again partitioned as in (B.1). The lower block of s_t consists of the elements of the reduced cognitive state $\bar{s}_t$, which is linear function of s_t since $\bar{s}_t = E[x_{t+1}|s_t]$. 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 (I.5) 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

$$\text{(C.2)} \qquad \mathrm{var}[\Lambda_{c,t}\underline{s}_t + \bar{\omega}_{t+1}] \;=\; (I - \Lambda_{d,t})X_t\Lambda'_{d,t},$$

where the matrix X_t defined as

$$X_t \;\equiv\; \mathrm{var}[\bar{s}_t]$$

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 (C.2) holds.)

We show that (1) forecast accuracy depends only on $\{\Lambda_{d,t}\}$, and (2) setting $\Lambda_{a,t} = \Lambda_{b,t} = \Lambda_{c,t} = 0$ is optimal, from which we conclude that optimal $\bar{m}_{t+1}$ is linear in $\bar{s}_t$ with an additive Gaussian noise.

1. *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 (C.2) is satisfied. Equation (C.2) 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 (C.2) 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 (III.2)), 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 (B.2), 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 (I.6) 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 (C.1) of the memory structure matter only to the extent that they have consequences for the information cost terms in (I.6).

2. Mutual information: a useful lemma

Information costs in period t are assumed to be an increasing function of $\mathcal{I}_t = \mathcal{I}\,(M; S)$, 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).[54] 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

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

(The information about the first set of variables that is revealed by learning the values of

[54] Here we adopt the notation used in Cover and Thomas (2006), with different symbols for the random variables M and S and their realizations. This is to make it clear that $\mathcal{I}_t$ 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; $\mathcal{I}_t$ is itself not a random variable, but a single number for each date t.

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

$$\begin{aligned}&\mathcal{I}\,(X_1,\ldots,X_k;\, X_{k+1},\ldots,X_{k+m}\,|X_{k+m+1},\ldots,X_{k+m+n})\\ &\quad\equiv\; H(X_1,X_2,\ldots,X_k\,|X_{k+m+1},\ldots,X_{k+m+n}) \;-\; H(X_1,\ldots,X_k\,|X_{k+1},\ldots,X_{k+m+n}).\end{aligned}$$

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

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

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

$$\mathcal{I}_t \;=\; \mathcal{I}\,(\underline{S},\bar{S};\,\underline{M},\bar{M}) \;\geq\; \mathcal{I}\,(\bar{S};\,\bar{M}). \tag{C.3}$$

Furthermore, this lower bound is achieved if and only if

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

For any three random variables X, Y, Z, the conditional mutual information $\mathcal{I}\,(X;\,Y\,|Z) = 0$ if and only if the variables X and Y are distributed independently one another, conditional on the value of Z. Hence the lower bound (C.3) 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.

3. 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 (C.1), 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 $\{\mathcal{I}_t\}$. 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 $\mathcal{I}_t$ given in (C.3) 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 (C.3) is achieved in each period. Thus such a specification achieves the lowest possible value for the combined objective function (I.6), 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 (I.6), but it will be essentially necessary. We have shown above that achieving the lower bound for $\mathcal{I}_t$ 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 (C.1) for the memory state with $\Lambda_{a,t} = \Lambda_{c,t} = 0$.[55] 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 $\mathcal{I}_t$ 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

[55] 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.

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 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 (I.6) 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 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

$$\text{(C.4)} \qquad \bar{m}_{t+1} \;=\; \bar{\Lambda}_t \bar{s}_t \;+\; \bar{\omega}_{t+1},$$

where $\bar{\Lambda}_t$ is the matrix called $\Lambda_{d,t}$ in (C.1). (This corresponds to equation (III.4) 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.

4. Properties that $\bar{\Lambda}_t$ should satisfy

In order for (C.4) to represent a memory structure consistent with the normalization according to which $\mathrm{E}[x_{t+1}\,|\bar{m}_{t+1}] = \bar{m}_{t+1}$, the sequence of matrices $\{\bar{\Lambda}_t\}$ and $\{\Sigma_{\bar{\omega},t+1}\}$ must satisfy certain properties. Note first that the condition (C.2) will be satisfied if and only if

$$\text{(C.5)} \qquad \Sigma_{\bar{\omega},t+1} \;=\; (I - \bar{\Lambda}_t) X_t \bar{\Lambda}_t'.$$

For $\Sigma_{\bar{\omega},t+1}$ to be a symmetric, p.s.d. matrix, the matrix $\bar{\Lambda}_t$ must satisfy the following properties: (a) the matrix $\bar{\Lambda}_t X_t = X_t \bar{\Lambda}_t'$ must be symmetric (so that the right-hand side of (C.5) is also symmetric); and (b) the right-hand side of (C.5) must be a p.s.d. matrix. 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. 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 (C.4) 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

$$\bar{\Lambda}_t \;=\; \lambda_t X_t v_t v_t', \tag{C.6}$$

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 (C.6) 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

$$\Sigma_{\bar{\omega},t+1} \;=\; \lambda_t (1 - \lambda_t) X_t v_t v_t' X_t. \tag{C.7}$$

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

$$\tilde{m}_{t+1} \;=\; \lambda_t v_t' \bar{s}_t \,+\, \tilde{\omega}_{t+1}, \qquad \tilde{\omega}_{t+1} \,\sim\, N(0,\, \lambda_t(1 - \lambda_t)). \tag{C.8}$$

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

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

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

$$\bar{\Lambda}_t \;=\; \lambda_t \frac{ww'}{w'w}, \tag{C.9}$$

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.[56] In this case, $\bar{\Lambda}_t$ is necessarily of the form (C.6), with the vector v_t given by

$$v_t \;=\; \frac{w}{(\Omega + \sigma_y^2)^{1/2}(w'w)}. \tag{C.10}$$

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 (C.10). In this special case, (C.7) reduces to

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

5. *An alternative representation for the reduced cognitive state*

Since $\bar{s}_t$ is defined as $E[x_{t+1}|s_t]$, we can decompose the variance of $var[x_{t+1}]$ as

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

from which we see that

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

Thus, the variance matrix of the reduced cognitive state $\bar{s}_t$ can be written as a function of the single parameter $\hat{\sigma}_t^2$. 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

$$\Gamma \;\equiv\; \begin{bmatrix} 1 & \frac{\Omega}{\Omega + \sigma_y^2} \\ 0 & 1 \end{bmatrix}. \tag{C.11}$$

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.

[56] 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 (III.7) in the main text also valid in the case that $X_t = X_0$, and thus it allows us to state certain conditions more compactly.

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

$$\text{var}[\check{s}_t] \;=\; \check{X}(\hat{\sigma}_t^2) \;\equiv\; \begin{bmatrix} \hat{\sigma}_0^2 - \hat{\sigma}_t^2 & 0 \\ 0 & \Omega + \sigma_y^2 \end{bmatrix}. \tag{C.12}$$

We can then alternatively write

$$X(\hat{\sigma}_t^2) \;=\; \Gamma\check{X}(\hat{\sigma}_t^2)\Gamma'. \tag{C.13}$$

D. The Law of Motion and the Information Content of Memory

We now consider how the parameterization of the law of motion (C.4) 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 $c\,(\mathcal{I}_t)$.

1. The degree of uncertainty implied by a given memory structure

We turn now to the question of how the memory-implied 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

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

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 (B.2), the left-hand side must equal the matrix Σ_0 defined there. Hence the variance decomposition can be written as

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

which implies that in any period,

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

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 (C.4) implies that

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

where the second line uses (C.5). Hence we obtain the prediction that

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

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 (D.1) into (B.5) 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 (D.1), together with (B.2), 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 (III.3) 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.

2. The mutual information implied by a given memory structure

Finally, we compute the mutual information $\mathcal{I}_t$ in the case that the memory state consists only of a reduced memory state $\bar{m}_{t+1}$, with law of motion (C.4). 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

$$D_{KL}(p(\cdot|x)||p(\cdot)) \equiv \mathrm{E}\left[\log \frac{p(y|x)}{p(y)}\right] \geq 0, \tag{D.2}$$

where the expectation is over possible realizations of y, and this quantity is a function of the particular realization x.[57] The *mutual information* $\mathcal{I}(X; Y)$ can then be defined as the mean value of this expression,

$$\mathcal{I}(X; Y) \equiv \mathrm{E}[D_{KL}(p(\cdot|x)||p(\cdot))], \tag{D.3}$$

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

where the expectation is now over possible realization of x, and the mutual information is also necessarily non-negative.[58]

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,

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

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 $\mathcal{I}(X;\, Y)$ will be the same as $\mathcal{I}(X;\, Z)$: it is unaffected by a change in the coordinates used to parameterize Y.[59]

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 (D.3) as above. Since mutual information should be independent of the coordinates used to parameterize the variables, we can use the value of $\mathcal{I}(X;\, Z)$ as our definition of $\mathcal{I}(X;\, Y)$ in this case as well (even though expression (D.2) 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 var$[Y|x]$ is of full rank. (Note that this matrix will be independent of the realization of x, and so can be written var$[Y|X]$, to emphasize that only the parameters of the joint distribution matter.) In this case var$[Y]$ is of full rank as well,

[58] 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.

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

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

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

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

$$\text{(D.4)} \qquad \mathcal{I}\,(X;\,Y) \;=\; -\frac{1}{2}\log\frac{\det(\mathrm{var}[Y|X])}{\det(\mathrm{var}[Y])}.$$

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

$$\text{(D.5)} \qquad \mathcal{I}\,(X;\,Y) \;=\; \mathcal{I}\,(X;\,Z) \;=\; -\frac{1}{2}\log\frac{\det(\mathrm{var}[Z|X])}{\det(\mathrm{var}[Z])}.$$

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 (C.4) 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 (D.4) to obtain

$$\text{(D.6)} \qquad \mathcal{I}_t \;=\; -\frac{1}{2}\log\frac{\det[(I-\bar{\Lambda}_t)X_t\bar{\Lambda}_t']}{\det[X_t\bar{\Lambda}_t']} \;=\; -\frac{1}{2}\log\det(I-\bar{\Lambda}_t),$$

in conformity with equation (III.7) 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 $\mathcal{I}_t$ implied by (D.6) 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 (C.6), as a consequence of which $\Sigma_{\bar{\omega},t+1}$ must be given by (C.7). 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 (C.8). 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 (D.5) to show that

$$\mathcal{I}_t \;=\; -\frac{1}{2}\log\frac{\lambda_t(1-\lambda_t)}{\lambda_t} \;=\; -\frac{1}{2}\log(1-\lambda_t), \tag{D.7}$$

Since in this case, $\det(I - \bar{\Lambda}_t) = \det(I - \lambda_t v_t v_t') = 1 - \lambda_t$, result (D.7) is again just what (D.6) would imply, so that (D.6) 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 (D.7) grows without bound. Thus we can assign the value $+\infty$ to $\mathcal{I}_t$ 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 (D.6) 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 (D.6) would imply, so that (D.6) 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 (C.6), with the vector v_t given by (C.10). 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 (C.8), and we can apply (D.5) to show that $\mathcal{I}_t$ will be given by (D.7). Again this is just what (D.6) would imply, so that (D.6) also yields the correct conclusion when X_t is a singular matrix.

Thus in all cases, (D.6) applies, and the value of $\mathcal{I}_t$ 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 $\{MSE_t, \mathcal{I}_t\}$, allowing the objective (I.6) to be evaluated.

E. 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

$$V(\hat{\sigma}_t^2) \;=\; \min_{\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))} [\alpha \hat{\sigma}_t^2 \;+\; c\,(\mathcal{I}\,(\bar{\Lambda}_t)) \;+\; \beta V(f(\hat{\sigma}_t^2, \lambda_t, v_t))]. \tag{E.1}$$

Here we establish some properties of the solution to the optimization problem on the right-hand side of (E.1) 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.

1. 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 (E.1), 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 (E.1)) 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 (E.1), and consider any two possible degrees of uncertainty $\hat{\sigma}_a^2, \hat{\sigma}_b^2$, satisfying

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

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 $\mathcal{I}\,(\bar{\Lambda}_b) < \infty$, 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

$$f(\hat{\sigma}_a^2, \bar{\Lambda}_a) \;=\; f(\hat{\sigma}_b^2, \bar{\Lambda}_b), \tag{E.3}$$

and in addition

$$\mathcal{I}\,(\bar{\Lambda}_a) \;\leq\; \mathcal{I}\,(\bar{\Lambda}_b). \tag{E.4}$$

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 (E.1).

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 (E.1)) 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 (E.1) 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 (E.2), the right-hand side of (E.1) 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 $0 \leq \lambda_b < 1$.[60] In this case, the memory structure for the following period is equivalent to one in which there is a univariate memory state

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

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

$$\Sigma_{t+1} \;=\; \Sigma_0 \;-\; \lambda_b(\Omega + \sigma_y^2) ww'. \tag{E.5}$$

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 (C.6), 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 (E.5). 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 (E.3). Moreover, for both

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

memory structures we have the same mutual information,

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

This establishes condition (E.4). Hence the value of the right-hand side of (E.1) 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 (E.1)) in this case be

$$\bar{m}_b \;=\; \bar{\Lambda}_b \bar{s}_b \,+\, \bar{\omega}_b, \tag{E.6}$$

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

$$\Sigma_{t+1} \;=\; \Sigma_0 \,-\, X_b\bar{\Lambda}_b'. \tag{E.7}$$

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

$$\bar{\Lambda}_a \;=\; \bar{\Lambda}_b\Gamma\Psi\Gamma^{-1}, \tag{E.8}$$

where Γ is the invertible matrix defined in (C.11), 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 (C.12). 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 (E.8) 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

$$\bar{m}_a \; = \; \bar{\Lambda}_a \bar{s}_a \; + \; \bar{\omega}_a, \tag{E.9}$$

where the transition matrix $\bar{\Lambda}_a$ is defined in (E.8), 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 (E.7); it follows that the implied value of $\hat{\sigma}_{t+1}^2$ is also the same as for the memory structure (E.6). 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 (E.9) involves no greater information cost than memory structure (E.6). 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, (D.6) implies that the mutual information will be given by

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

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 (E.4) 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 (E.8) implies that $\bar{\Lambda}_a = 0$ as well. In this case, $\det(I-\bar{\Lambda}_a) = \det(I-\bar{\Lambda}_b) = 1$, so that $\mathcal{I}(\bar{\Lambda}_a) = \mathcal{I}(\bar{\Lambda}_b) = 0$, and (E.4) 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 $0 < \lambda_b < 1$[61] 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 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 (E.8) implies that the columns of $\bar{\Lambda}_a$ must also all be multiples of the vector X_bv_b. It follows that $\bar{\Lambda}_a$ must also be singular, and that its unique non-null eigenvector must be

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

$X_b v_b$, 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 (E.4) must hold in this case as well.

Thus we have shown that whenever $\hat{\sigma}_a^2, \hat{\sigma}_b^2$ satisfy (E.2), 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 (E.3) and (E.4). This means that it must be possible to achieve a lower value for the right-hand side of (E.1) 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 (E.1) 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 (E.1) must be a monotonically increasing function of its argument.

2. *Optimality of a unidimensional memory state*

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

$$\min_{\bar{\Lambda}_t \in \mathcal{L}(X(\hat{\sigma}_t^2))} \mathcal{I}\,(\bar{\Lambda}_t) \qquad \text{s.t. } f(\hat{\sigma}_t^2, \bar{\Lambda}_t) \;\leq\; \hat{\sigma}_{t+1}^2, \tag{E.10}$$

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 (E.10), 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 $\mathcal{I}\,(\bar{\Lambda}_t)$.

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.

It is useful to observe that in any period t, the Kalman filter (III.1) 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

$$\hat{\mu}_t = \psi_t + \delta' \bar{m}_t. \tag{E.11}$$

where $\delta_{t+1} \equiv e_1 \; \gamma_{1,t+1} c$.

Then let the alternative transition matrix be given by

$$\bar{\Lambda}_t^{1D} \;=\; \lambda_t X_t v_t v_t', \tag{E.12}$$

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 we let the matrix $\Sigma_{\bar{\omega},t+1}$ be correspondingly modified, i.e. $\Sigma_{\bar{\omega},t+1} = (I - \bar{\Lambda}_t^{1D})X_t\bar{\Lambda}_t^{1D\prime}$. 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

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

where the final inequality follows from the fact that $\bar{\Lambda}_t X_t \bar{\Lambda}_t'$ 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 $v_t' X_t v_t = 1$. 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 $\bar{m}_{t+1} = X_t v_t \tilde{m}_{t+1}$, where $\tilde{m}_{t+1}$ is the unidimensional memory state with law of motion (III.13). From this it follows that

$$\delta_{t+1}'\bar{m}_{t+1} \;=\; \lambda_t \delta_{t+1}' X_t v_t v_t' \bar{s}_t \;+\; \delta_{t+1}' X_t v_t \tilde{\omega}_{t+1}$$

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

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

and

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

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 (E.11)), 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

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

will be the same. Thus $\bar{\Lambda}^{1D}_t$ also satisfies the constraint in (E.10).

Next we show that $\mathcal{I}(\bar{\Lambda}^{1D}_t)$ must be lower than $\mathcal{I}(\bar{\Lambda}_t)$. 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 (E.11) 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 (D.6) and (D.7) then imply that $\mathcal{I}\,(\bar{\Lambda}_t^{1D}) < \mathcal{I}\,(\bar{\Lambda}_t)$.

Thus $\bar{\Lambda}_t$ cannot be the solution to the optimization problem (E.10). 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.

3. The optimal weight vector of the 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 (C.8). 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

$$\hat{\mu}_{t+1} \ = \ \xi\tilde{m}_{t+1} \ + \ \gamma_1 y_{t+1} \qquad \text{(E.13)}$$

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

$$f(\hat{\sigma}_t^2, \lambda_t;\, \psi, \gamma_1) \;=\; \mathrm{var}[\mu \;-\; \psi' \bar{s}_t \;-\; \gamma_1 y_{t+1}] \;+\; \frac{1-\lambda_t}{\lambda_t} \psi' X_t \psi. \tag{E.14}$$

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 (E.14), 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 (E.14) are given by the linear equations

$$\psi \;=\; \lambda_t (e_1 - \gamma_1 c), \tag{E.15}$$

$$c' X_t (e_1 - \gamma_1 c - \psi) \;+\; (1-\rho)(1 - (1-\rho)\gamma_1)\hat{\sigma}_t^2 \;-\; \gamma_1 \sigma_\epsilon^2 \;=\; 0. \tag{E.16}$$

Equation (E.15) already allows one valuable insight: the optimal weight vector v_t is simply a normalized version of the vector δ_{t+1} defined in (E.11). 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 (E.15)–(E.16) 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

$$\gamma_{1,t+1} \;=\; \frac{(1-\lambda_t)\Omega + \lambda_t (1-\rho)\hat{\sigma}_t^2}{(1-\lambda_t)(\Omega + \rho^2 \sigma_y^2) + \lambda_t (1-\rho)^2 \hat{\sigma}_t^2 + \sigma_\epsilon^2} \tag{E.17}$$

as an explicit solution for the Kalman gain. It is worth noting that this implies that

$$0 \;<\; \gamma_{1,t+1} \;<\; \frac{1}{1-\rho}. \tag{E.18}$$

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 (E.15). Substituting the solutions for γ_1 and ψ into the quadratic objective, we obtain for the minimum possible value of the objective

$$\hat{\sigma}_{t+1}^2 \;=\; (1-\lambda_t)\delta'_{t+1}\Sigma_0\delta_{t+1} + \lambda_t(\delta_{1,t+1})^2\hat{\sigma}_t^2 \;+\; \gamma_{1,t+1}^2\sigma_\epsilon^2. \tag{E.19}$$

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}.

F. The Simple Example in Section II

Posterior uncertainty about the long-run mean μ sequentially evolves according to (II.3), (II.6) and (II.7). It is straightforward to see that the posterior uncertainty converges to a limit after an extensive learning, i.e. $\Sigma_t \to \Sigma_\infty$ and $\hat{\sigma}_t^2 \to \hat{\sigma}_\infty^2$. From the equations, one can derive that Σ_∞ should satisfy

$$\Sigma_\infty = (1-\bar{\lambda})\,\Omega + \bar{\lambda}\left(\frac{1}{\Sigma_\infty} + \frac{1}{\sigma_y^2}\right)^{-1}.$$

Rearranging this term yields a unique solution for Σ_∞ as follows,

$$\Sigma_\infty = \frac{\sigma_y^2}{2}\left\{-(1-\bar{\lambda})\left(1-\frac{\Omega}{\sigma_y^2}\right) + \sqrt{(1-\bar{\lambda})^2\left(1-\frac{\Omega}{\sigma_y^2}\right)^2 + 4(1-\bar{\lambda})\frac{\Omega}{\sigma_y^2}}\right\}.$$

Thus, in the perfect memory case ($\bar{\lambda} = 1$), the posterior uncertainty converges to zero, $\Sigma_\infty = \hat{\sigma}_\infty^2 = 0$. In comparison, for $\bar{\lambda} < 1$, we have $\Sigma_\infty > 0$ and $\sigma_\infty^2 > 0$. Then, (II.2) determines the long-run Kalman gain as

$$\gamma_\infty = \frac{\Sigma_\infty}{\Sigma_\infty + \sigma_y^2},$$

which is positive as long as $\bar{\lambda} < 1$.

G. Numerical Solutions

Here we provide further details of the numerical calculations reported in section V of the main text.

1. 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 (E.19) becomes

$$\eta_{t+1} \;=\; (1-\lambda_t)(1-\gamma_{1,t+1})^2 K \;+\; (1-\rho^2\lambda_t)\gamma_{1,t+1}^2 \;+\; \lambda_t(1-(1-\rho)\gamma_{1,t+1})^2\eta_t, \tag{G.1}$$

and (E.17) becomes

$$\gamma_{1,t+1} \;=\; \frac{(1-\lambda_t)K + (1-\rho)\lambda_t\eta_t}{(1-\lambda_t)(K+\rho^2) + (1-\rho^2) + (1-\rho)^2\lambda_t\eta_t}. \tag{G.2}$$

Substitution of (G.2) for $\gamma_{1,t+1}$ in the right-hand side of (G.1) 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 (E.14) 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.

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 A1 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 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 A2 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.

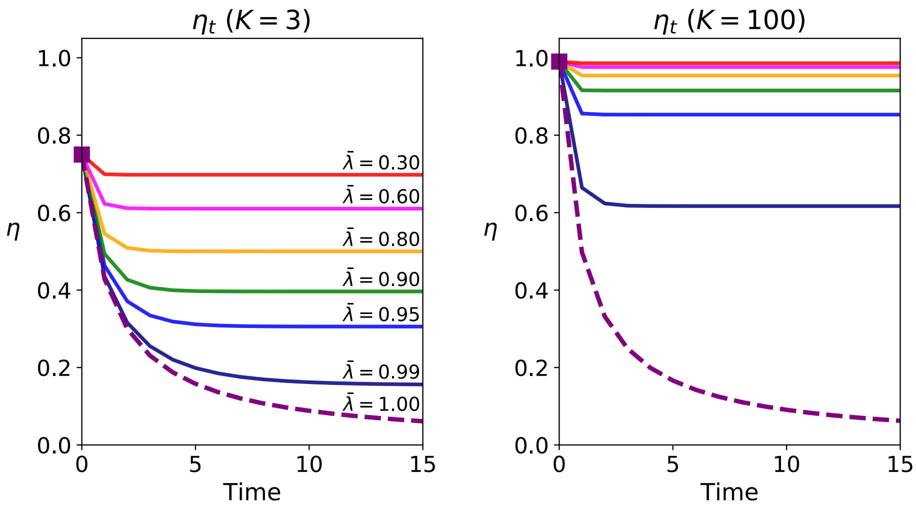

Figure A1. The evolution of uncertainty about μ (when $\rho = 0$)

Note: 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).

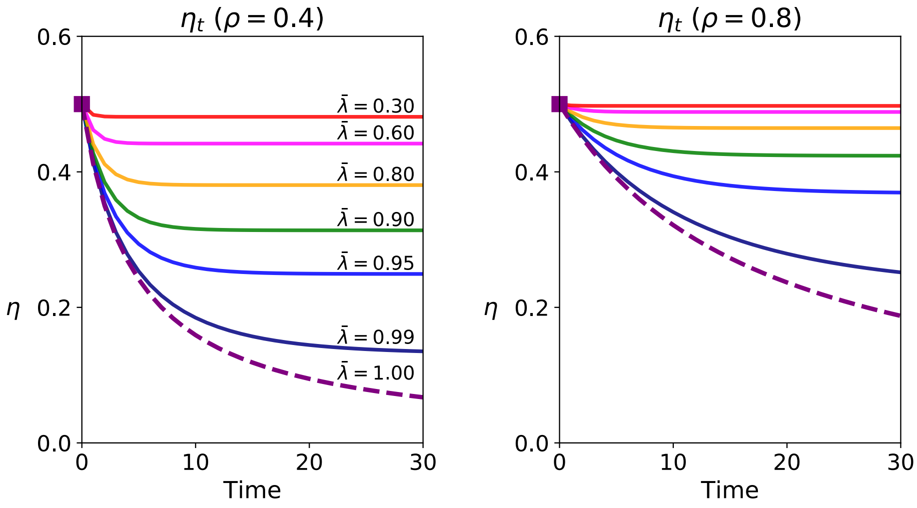

Figure A2. The evolution of uncertainty about μ (when $\rho > 0$)

Note: 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).

2. *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 $\mathcal{I}_t$), 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 IV.B).

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

$$\text{(G.3)} \qquad \tilde{V}(\eta_t) = \min_{\lambda_t \in [0,1]} \left[\eta_t - \frac{\tilde{\theta}}{2} \log(1 - \lambda_t) + \beta \tilde{V}(\phi(\eta_t; \lambda_t)) \right].$$

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 A5 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 (G.3).

This function is graphed for several values of $\tilde{\theta}$ in Figure A3, 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 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

$$\text{(G.4)} \qquad \eta_{t+1} = \phi(\eta_t; \lambda_t)$$

derived in section G.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 by these equations can be graphed in a phase diagram, as illustrated in Figure A4. 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 (G.4), 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

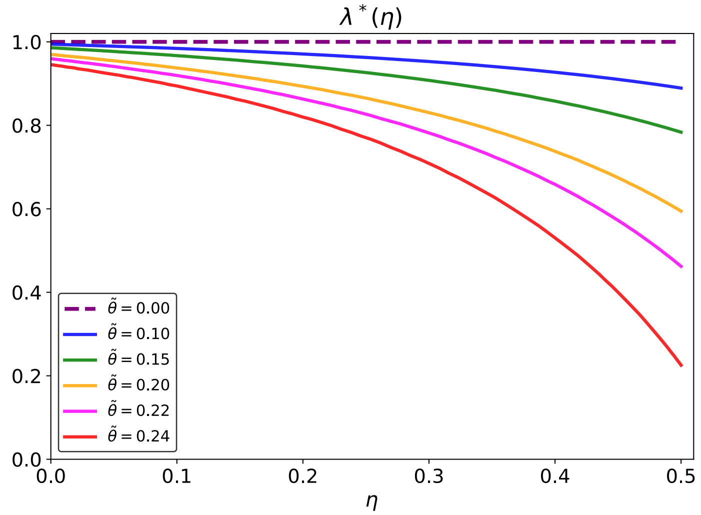

FIGURE A3. THE OPTIMAL POLICY FUNCTION

Note: The optimal policy function $\lambda^*(\eta)$, in the case of progressively larger values for the information cost parameter $\tilde{\theta}$, under the assumption that $K = 1, \rho = 0$.

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 (G.4) 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.[62] 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 A3.[63] Again the values of (η_t, λ_t) in successive periods must always lie on the graph of the policy function; the direction of

[62] 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.

[63] 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 A3. 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 (G.4) remain the same as in the left panel.

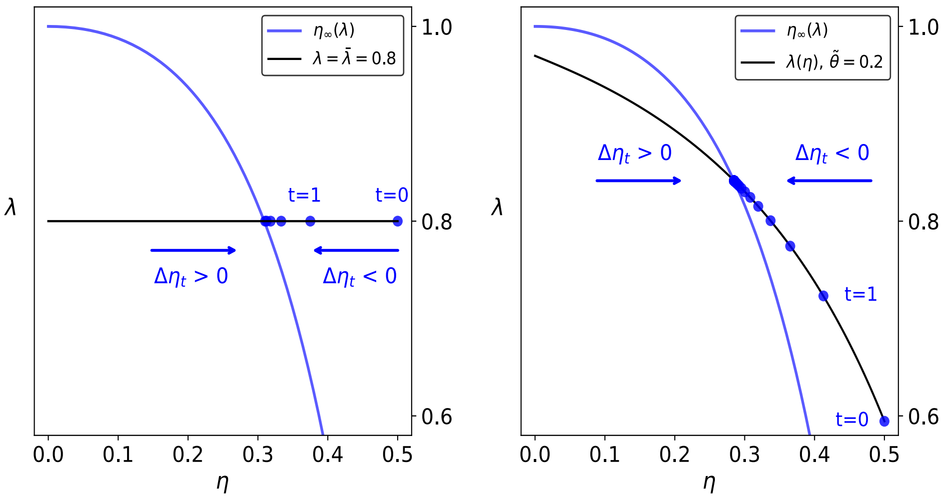

FIGURE A4. THE DYNAMICS OF SCALED UNCERTAINTY AND MEMORY PRECISION

Note: 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 $\tilde{\theta}$.

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 A3, and hence to different dynamics in the phase plane, converging to 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.

3. *The possibility of discontinuous solutions*

Figure A5 illustrates our comment about the possible non-convexity of the optimization problem (G.3). Let $RHS(\lambda_t; \eta_t)$ be the function defined on the right-hand side of (G.3), 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

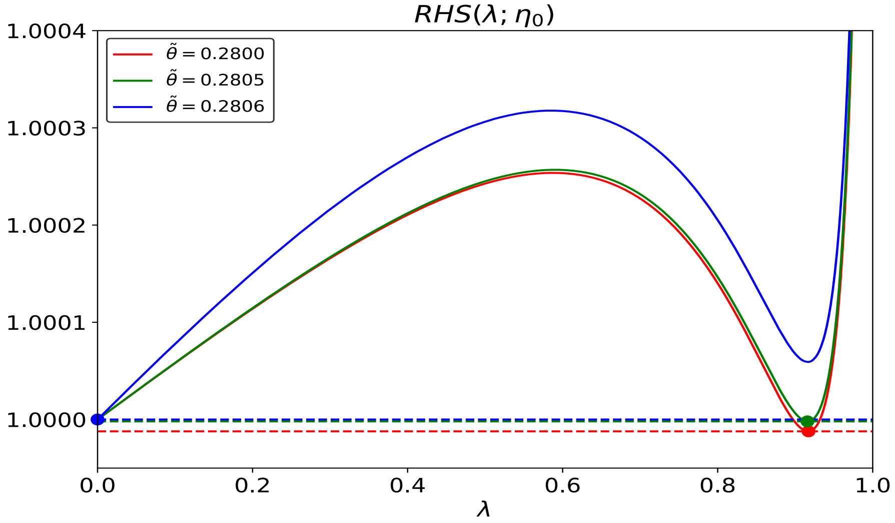

Figure A5. The Bellman equation

Note: 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.

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 $\phi\left(\eta_0;\lambda_t\right)$ 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 $\phi\left(\eta_0;\lambda\right)$ to the value of λ; see equation (G.5) 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 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 A3, the optimal policy function $\lambda^*(\eta)$ is equal to zero for all large enough (though still finite) values of η, as illustrated in Figure A6. 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.

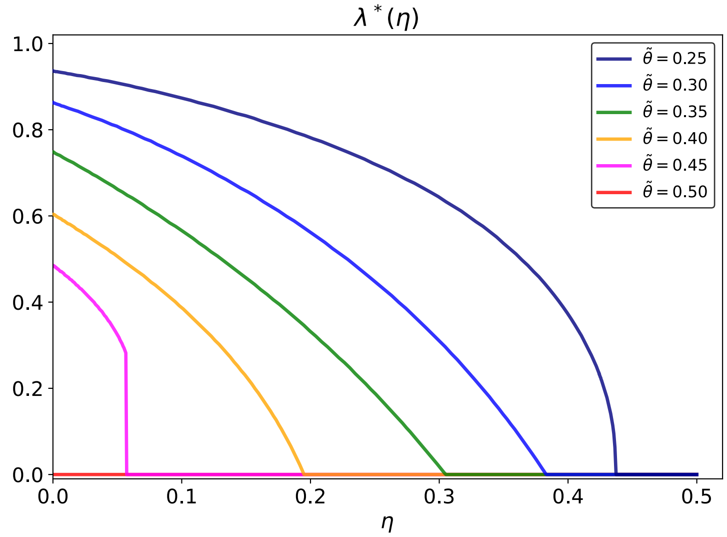

FIGURE A6. THE OPTIMAL POLICY FUNCTION (FOR A SUFFICIENTLY LARGE $\tilde{\theta}$)

Note: 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 A3.

4. 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 G.1) simplifies to

$$\eta_{t+1} = 1 - \frac{1}{K + 1 - \lambda_t (K - \eta_t)} \;\equiv\; \phi\left(\eta_t; \lambda_t\right). \tag{G.5}$$

In the case of an exogenous upper bound on mutual information, the nonlinear difference equation obtained by setting $\lambda_t = \bar{\lambda}$ in (G.5) 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 $\phi\left(\eta_t; \bar{\lambda}\right)$ 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 (G.5) 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

$$\tilde{V}(\eta_t) = \min_{\lambda_t} \left[\beta^2 \eta_t - \frac{\tilde{\theta}}{2} \log\left(1 - \lambda\right) + \beta \tilde{V}\left(\phi\left(\eta_t; \lambda_t\right)\right) \right].$$

The first order condition with respect to λ_t is

$$\frac{\tilde{\theta}}{2} \frac{1}{1 - \lambda_t} + \beta \tilde{V}'(\eta_{t+1}) \frac{\partial \phi\left(\eta_t; \lambda_t\right)}{\partial \lambda_t} = 0. \tag{G.6}$$

And the envelope condition is

$$\tilde{V}'(\eta_t) = \beta^2 + \beta\tilde{V}'(\eta_{t+1})\frac{\partial\phi\left(\eta_t;\lambda_t\right)}{\partial\eta_t}. \tag{G.7}$$

We can use these two conditions to derive an Euler equation for the dynamics of the scaled uncertainty measure.

Substituting the solution (G.5) for $\phi\left(\eta_t;\lambda_t\right)$ and taking the derivative with respect to λ_t, we can rewrite (G.6) as

$$\begin{aligned}\tilde{V}'(\eta_{t+1}) &= -\frac{\tilde{\theta}}{2\beta}\frac{1}{1-\lambda_t}\left(\frac{\partial\phi\left(\eta_t;\lambda_t\right)}{\partial\lambda_t}\right)^{-1}\\ &= -\frac{\tilde{\theta}}{2\beta}\frac{1}{1-\lambda_t}\left(-\frac{(K-\eta_t)}{\left(K+1-\lambda_t(K-\eta_t)\right)^2}\right)^{-1}\\ &= \frac{\tilde{\theta}}{2\beta}\frac{\left(K+1-\lambda_t(K-\eta_t)\right)^2}{(1-\lambda_t)(K-\eta_t)}\\ &= \frac{\tilde{\theta}}{2\beta}\frac{1}{\left(1-\eta_{t+1}\right)\left(1-(1-\eta_{t+1})(1+\eta_t)\right)},\end{aligned}$$

where the last equality is derived by again substituting the law of motion (G.5). It follows that if $\eta_t \to \eta_\infty$ in the long run, the stationary solution η_∞ must satisfy

$$\tilde{V}'(\eta_\infty) = \frac{\tilde{\theta}}{2\beta}\frac{1}{(1-\eta_\infty)\eta_\infty^2}. \tag{G.8}$$

Next we rewrite (G.7), again taking the derivative of expression (G.5) for $\tilde{V}(\eta_t;\lambda_t)$:

$$\begin{aligned}\tilde{V}'(\eta_t) &= \beta^2 + \beta\tilde{V}'(\eta_{t+1})\frac{\partial\phi\left(\eta_t;\lambda_t\right)}{\partial\eta_t}\\ &= \beta^2 + \beta\tilde{V}'(\eta_{t+1})\frac{\lambda_t}{\left(K+1-\lambda(K-\eta_t)\right)^2}\\ &= \beta^2 + \beta\tilde{V}'(\eta_{t+1})\frac{\lambda_t}{\left(1-\eta_{t+1}\right)^{-2}}\\ &= \beta^2 + \beta\tilde{V}'(\eta_{t+1})(1-\eta_{t+1})^2\frac{(K+1)(1-\eta_{t+1})-1}{(K-\eta_t)(1-\eta_{t+1})}.\end{aligned}$$

It follows that the stationary solution η_∞ must satisfy

$$\tilde{V}'(\eta_\infty) = \beta^2 + \beta\tilde{V}'(\eta_\infty)\frac{(1-\eta_\infty)\left[(K+1)(1-\eta_\infty)-1\right]}{K-\eta_\infty}. \tag{G.9}$$

Moreover, in a stationary solution, the value $\tilde{V}'(\eta_\infty)$ given by (G.8) must also be the value of $\tilde{V}'(\eta_\infty)$ in (G.9). Using (G.8) to substitute for $\tilde{V}'(\eta_\infty)$ in (G.9), 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 $0 < \eta_\infty < K/(K+1)$):

$$\tilde{\theta} = 2\beta^3(1-\eta_\infty)\eta_\infty^2\left[1-\beta\frac{(K+1)(1-\eta_\infty)^2-(1-\eta_\infty)}{K-\eta_\infty}\right]^{-1}. \tag{G.10}$$

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 $\tilde{\theta}$.

H. Predicted Values for the Quantitative Measures of Forecast Bias

Here we provide further explanation of the numerical results reported in section V of the main text.

1. 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 $\rho_m \equiv \lambda_\infty (e_1' v_\infty)\,(e_1' - \gamma_1 c')\, X_\infty v_\infty$ and $\rho_{my} \equiv \lambda_\infty\,(\gamma_1 + e_2' v_\infty)$.

We can evaluate the numerical values of the coefficients defining the long-run dynamics as follows. Equations (G.1)–(G.2) 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

$$0 \;<\; \rho_m \;<\; 1. \tag{H.1}$$

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.1) 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 IV.C. 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 A7 and A8 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 $K = 1$.)

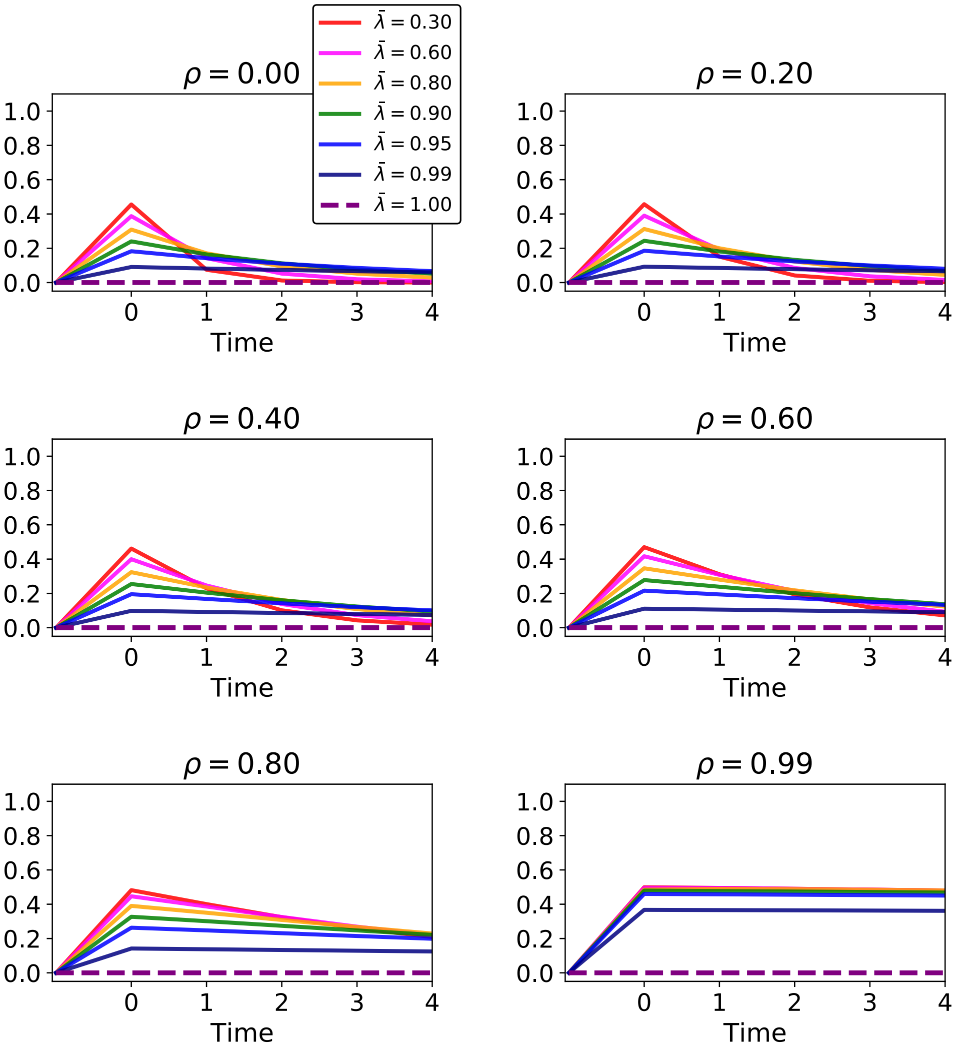

Figure A7. Impulse responses of the DM's estimate of μ for alternative degrees of persistence ρ of the external state process.

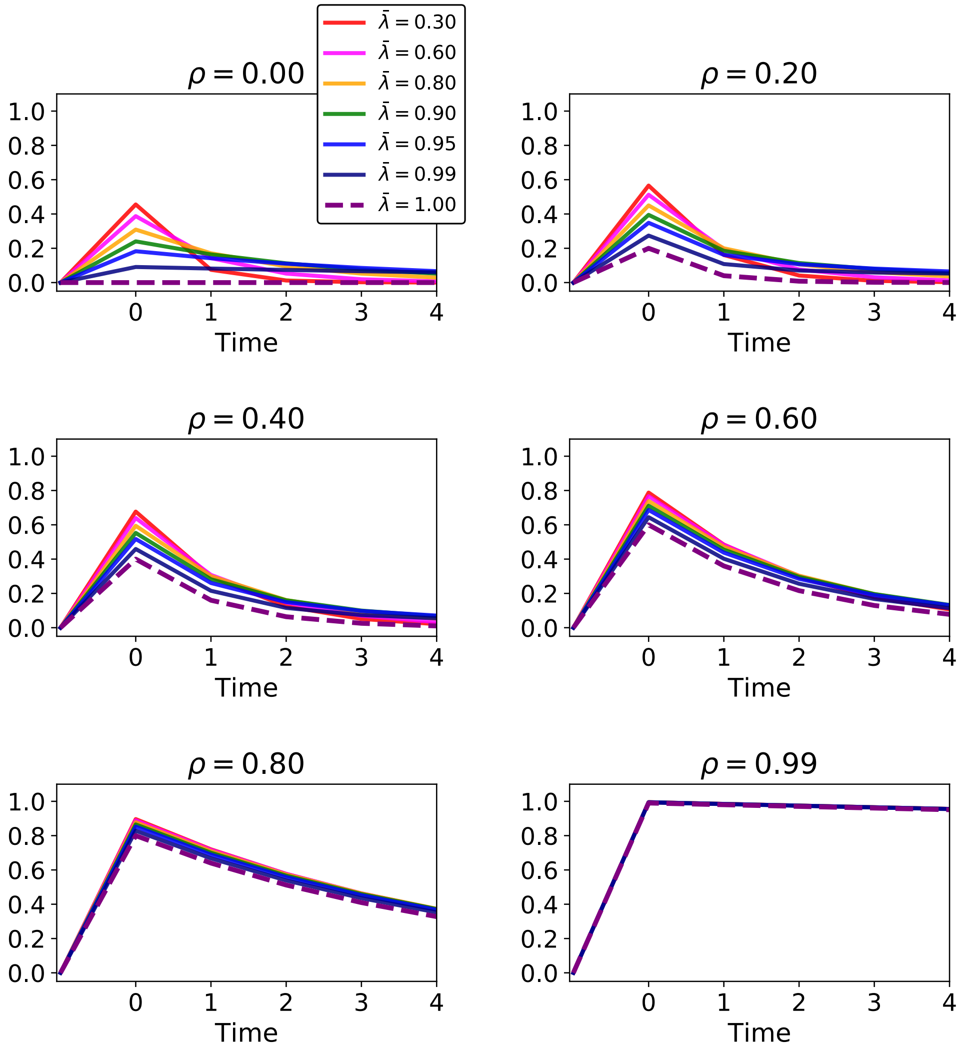

Figure A8. Impulse responses of the DM's one-quarter-ahead forecast of the external state for alternative degrees of persistence ρ of the external state process.

2. Predicted value of the regression coefficient $b_{s,h}$

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

$$b_{s,h} \equiv \frac{cov[\hat{y}_{t+h|t}, y_t|\mu]}{var[y_t|\mu]}$$

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

$$\begin{aligned} b_{s,h} &= (1-\rho^h)\frac{cov[\hat{\mu}_t, y_t|\mu]}{var[y_t|\mu]} + \rho^h \\ &= (1-\rho^h)\left[(e_1' - \gamma_1 c')\left(\tilde{\Sigma}_{\bar{m}} - \frac{1}{K}\tilde{\Sigma}_{\bar{m}} e_1 e_1' \tilde{\Sigma}_{\bar{m}}\right) c + \gamma_1\right] + \rho^h. \end{aligned}$$

In Figure 7, the value of $b_{s,h}^{\frac{1}{h}}$ is plotted against the value of ρ.

3. *Parameterization of the Model*

We find pairs of parameters $(\bar{\lambda}, K)$ that minimize the following target.

$$MSE^{targeted} = \frac{1}{6}\sum_{\rho}\left(\frac{\rho_1^s - \hat{\rho}_1^s}{\hat{\rho}_1^s}\right)^2$$

where $\hat{\rho}_1^s$ is the degree of over-reactions observed in one-period-ahead forecasts data, and ρ takes values from $[0.0, 0.2, 0.6, 0.8, 1.0]$. The minimum sample MSE is achieved at 0.0011, and the best-fitting pairs are displayed in the left panel of Figure A9. The right panel displays the pairs of $(\bar{\lambda}, K)$ that generate the same level of $MSE^{targeted}$. We can see that each curve is upward-sloping. This is because a lower degree of over-reactions predicted by a higher $\bar{\lambda}$ has to be offset by a higher K.

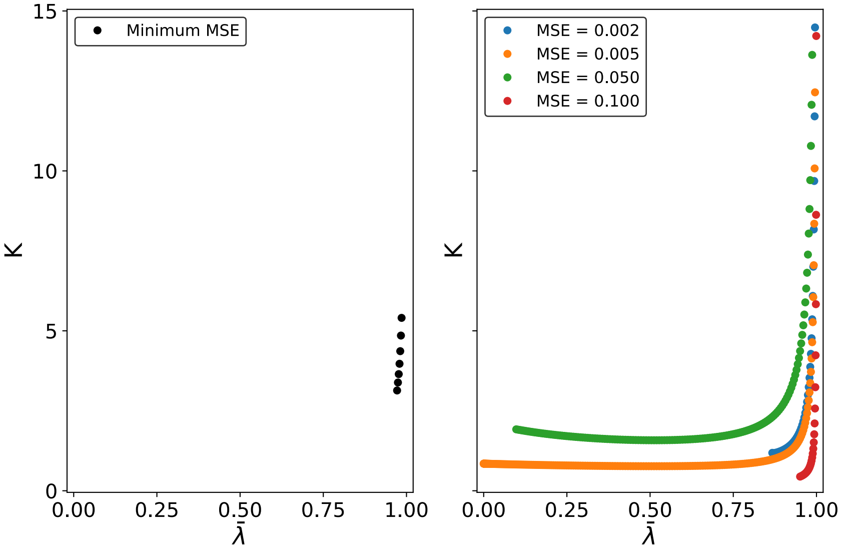

FIGURE A9. PAIRS OF $(\bar{\lambda}, K)$ GENERATING THE SAME LEVEL OF $MSE^{targeted}$