Learning and Subjective Expectation Formation: A Recurrent Neural Network Approach *

Chenyu Hou †

First Draft: August, 2020
This Draft: April, 2023

Abstract

Most empirical studies on expectation formation models share a common dynamic structure but impose different functional form restrictions. I propose a flexible non-parametric method that maintains this dynamic structure to estimate a model of expectation formation using Recurrent Neural Networks. Applying this approach to data on macroeconomic expectations from the Michigan Survey of Consumers and a rich set of signals, I document three novel findings: (1) agents' expectations about the future economic condition have asymmetric and non-linear responses to signals; (2) agents' attentions shift from signals about the current state to signals about the future as the economic condition deteriorates ; (3) the content of signals on economic conditions plays the most important role in creating the attention-shift. Double Machine Learning approach is used to obtain statistical inferences of these empirical findings. Finally, I show these stylized facts can be generated by a model with rational inattention, in which information endogenously becomes more valuable when economic status worsens.

Keywords: Expectation Formation, Bounded Rationality, Information Acquisition, Non-parametric Method, Recurrent Neural Network, Survey Data

*I'm grateful to Jesse Perla, Paul Schrimpf, Paul Beaudry, Michael Devereux, and Amartya Lahiri for their invaluable guidance and support on this project. I thank Vadim Marmer, Fabio Milani, Monika Piazzesi, Henry Siu, Hassan Afrouzi, and many others for their insightful comments. I'm also thankful to the computational support of Compute Canada (www.computecanada.ca). All the remaining errors are mine.

†Chenyu Hou: Vancouver School of Economics, University of British Columbia. Email: sevhou1989@gmail.com

1 Introduction

Models of expectation formation have played an important role in modern macroeconomic theories. The past decade has seen a surge of empirical studies using survey data to examine how information about aggregate economic status, such as unemployment and inflation rate, affects households' macroeconomic expectations. For example, in their seminal work, Coibion and Gorodnichenko (2012) document pervasive evidence that expectations from the Michigan Survey of Consumers (MSC) deviate from Full Information Rational Expectation (FIRE) and conclude that households have limited information. However, these empirical frameworks usually use restrictive assumptions on functional forms to apply parametric methods. As a result, empirical findings with these approaches are subject to these parametric assumptions and might miss important features of the relationship between households' macroeconomic expectations and signals. For example, when facing information about different macroeconomic aspects or from various sources, agents may be selective about the information they use to form expectations. Positive and negative news about economic status may have different impacts in terms of magnitude on their expectations. Furthermore, the way they utilize various information may differ when the state of the economy changes.[1] This paper aims to explore whether these patterns exist in the data.

To achieve this goal, I first make a methodological contribution by proposing an empirical framework that allows for a flexible relationship between macroeconomic signals and households' expectations. I show that most expectation formation models in macroeconomics adopt a common dynamic structure, where households form expectations about the future by perceiving some latent variables according to some signals. However, the relations between signals, latent variables, and expectations take different forms depending on the parametric assumptions made in the model. For example, in the standard noisy information model, the latent variable is the posterior mean of that state and has a linear relation with expectations. In Markov Switching Models, these latent variables become their posterior beliefs on the Markovian state, and their relation with expectations is governed by Bayes Rule.

The novelty of my empirical method is that I impose no restrictions on what the latent variables are, how the signals affect the latent variables, and how the latent variables affect households' expectations. Meanwhile, proper restrictions are made to maintain the dynamic structure described above. The relation between signals and expectational variables through the dynamic structure is estimated using a non-parametric method, Recurrent Neural Networks (henceforth RNN). RNN can be used in this specific context because it can universally approximate the dynamic system that represents the general structure proposed above.[2] This

[1]For example, Coibion and Gorodnichenko (2015) documents that the level of information rigidity falls in recessions and is particularly high during the Great Moderation. This indicates that the way economic agents process information may change as economic status changes.

[2]See Schäfer and Zimmermann (2006) for the universal approximation property of RNN in the context of

method offers a way to capture the flexible relationship between signal and expectational variables without making further parametric assumptions on functional forms except for the common dynamic structure. In particular, suppose the macroeconomic signals affect the expectations non-linearly, or through interacting with other signals or the latent variables. In these cases, the relations will be captured by RNN but are usually missed by models that are linear or with pre-assumed structures. On the other hand, if the underlying mapping between signals and expectations is linear, this approach will uncover a linear relationship.

The estimated functional form offers important insight on plausible structures for households' expectation formation process. It also provides a way to evaluate how macroeconomic signals affect households' expectations. Following the functional estimation with RNN, I apply the Double Machine Learning (DML) method proposed by Chernozhukov *et al.* (2018) to estimate the average marginal effect of the macroeconomic signals on households' expectations. This approach is usually used to correct the bias induced by the plug-in estimators following machine learning methods. It is also known to deliver valid inferences on these estimators under high-level assumptions on the corresponding moment condition model and machine learning estimators, thus allowing for tests on the statistical significance of my empirical findings.

Applying my empirical methods to the Michigan Survey of Consumers, I document three major findings new to the literature. I first show that households' expectations of the economic conditions, namely the unemployment rate and the real GDP growth, are *non-linear* functions of signals about the change of unemployment rate and real GDP - the effect of an incremental change in such a signal depends on the level of the signal itself. The relationship is also *asymmetric* - positive and negative signals with the same magnitude have an asymmetric impact on expectations. In particular, households respond more aggressively to signals that suggest the economic status worsens.

Furthermore, I find the marginal effects of these signals change over time. The absolute values for the marginal effects of signals on the economic conditions fall as the GDP growth slows down or the unemployment rate hikes up. However, the opposite is true for the signals that contain information about the future. When interpreting marginal effects as weights that households put on signals, this finding suggests that households *shift their attention* from signals about current and past states to those about the future. In other words, the households behave as "adaptive learners" when economic conditions are stable and become more "forward-looking" when the situation gets worse.

Lastly, the estimated functional form of the expectation formation model suggests such an attention-shift is mainly driven by the signals on economic conditions rather than information related to the interest rate or inflation. Furthermore, they contribute to the attention-shift through both the contemporaneous signals newly observed in each period and the latent vari-

a dynamic system.

ables that capture the past signals' impacts. This is consistent with the empirical evidence on the presence of info rigidity largely documented in the literature on households' expectations. Moreover, in my empirical framework, I also include measures on the amount of news coverage about various macroeconomic aspects from both local and national news media.[3] I refer to such a measure as "volume of news". I then find that a higher volume of news about the economic condition from media leads to a higher weight on signals about the future, as suggested by Carroll (2003). However, it does not explain the drop of weights on signals about current and past states. Instead, it is the *content* of signals on economic conditions, rather than the *volume* of news on these signals, that plays the most important role in creating the attention-shift.

These new stylized facts are consistent with rational inattention models but hard to be reconciled with many other frameworks for modeling beliefs. For example, for a model with Full Information Rational Expectation to explain the attention-shift between signals on current and future states, one has to believe that the economic conditions, such as the unemployment rate, follow a more persistent or volatile process during recession episodes. Standard noisy information and sticky information models are also insufficient. To create weight changes on signals in these models, one needs state-dependency in either precision of signals or the underlying state-space model that agents believe in, both of which are exogenous in those models. One possible explanation for the attention-shift is through the volume of news reported by media as first proposed in Carroll (2003). Lamla and Lein (2014) formalized the idea by showing that greater media coverage increases the precision of signals about the future in agents' signal-extraction problem, leading to higher weights on these signals. For this explanation to work, one should observe that the weights on the current signals fall as the volume of news on economic conditions increases. Moreover, the volume of this news alone should account for most of the variations in the change of marginal effects. However, neither of these is true according to my empirical findings.

I then develop a simple model featuring rational inattention to explain these stylized facts. When agents have limited ability to acquire information, they will choose to allocate their limited resources optimally on a subset of signals available to them. These choices can change as economic status changes, thus creating the attention-shift and the non-linear responses to different signals. Moreover, the state-dependency created by this type of model is not ad hoc: it comes from agents' optimal behavior in the face of information constraints. In the rational inattention model I propose, information about the future becomes more valuable endogenously when the state of the economy gets worse. For this reason, households start to seek more information about the future actively and end up placing higher weights on these signals when forming their expectations.

[3] I scraped the number of news stories on related macroeconomic topics (i.e. inflation, interest rate and unemployment rate) from TV news scripts and local newspaper articles in LexisNexis Database. Then I construct a measure of news coverage on these topics following PFAJFAR and SANTORO (2013).

Literature Review This paper contributes to several different strands of literature. It first relates to the growing literature using machine learning techniques to solve macroeconomic problems. There is a surge in applications of modern machine learning tools in economics for the past several years, including prediction problems as discussed in Kleinberg *et al.* (2015) as well as more recent work on causal inference such as in Athey and Imbens (2016) and Chernozhukov *et al.* (2017).[4] The empirical method of this paper is closely related to those in Chernozhukov *et al.* (2018) and Farrell *et al.* (2021). The latter offers convergence-speed conditions for deep Neural Networks to acquire valid inference. The average marginal effect derived in this paper is a form of the *average derivative* described in Chernozhukov *et al.* (2022). Another paper closely related to this is Bianchi *et al.* (2022). The authors use Elastic Net to form benchmark macroeconomic forecasts in a data-rich environment and use them to assess possible distortions in survey expectation data. My paper focuses on the estimation of possibly sub-optimal weights on information used by households when forming expectations. The estimated results are used later to shed light on how to model the expectation formation process. To the best of my knowledge, this is the first time RNN is applied to learning and expectation formation problems in an estimation context.

This paper also relates to the growing empirical literature using survey data to investigate how expectations are formed. These studies have documented substantial evidence that agents' expectations are formed under a limited information structure (Coibion and Gorodnichenko (2012), Andrade and Le Bihan (2013), Lamla and Lein (2014) etc), using various sources of information (Carroll (2003), Lamla and Lein (2014), D'Acunto *et al.* (2020) etc). Whereas this paper focuses on the non-linear, asymmetric, and state-dependent responses of expectations to macroeconomic signals. Related to this matter, a recent paper Roth *et al.* (2020) finds that U.S. households demand an expert forecast about the likelihood of recession when perceiving higher unemployment risk in a random experiment setting. My paper adds to this new literature using observational data by showing that various sources of information compete for households' attention, and they acquire more information about the future from experts when the state of the economy gets worse.

The dynamic structure in my empirical framework is built on the literature about learning and information acquisition. This literature has a long history in macroeconomics. The models developed in this literature include Constant Gain Learning (e.g. Evans and Honkapohja (2001), Milani (2007), Eusepi and Preston (2011)),[5] Noisy Information (e.g. Woodford (2001)), Markov Regime Switching (e.g. Hamilton (2016)) and Rational Inattention (e.g. Sims (2003), Mackowiak and Wiederholt (2009), Maćkowiak *et al.* (2018)). All these models adopt the same dynamic structure as in my empirical framework but differ in parametric functional

[4]For a complete review on recent applications of Machine Learning tools in economics, see Athey (2018).

[5]The Constant Gain Learning Framework is later extended to models in which experiences affect expectations (Malmendier and Nagel (2015)), and models to explain heterogeneity across agents (Cole and Milani (2020)).

form assumptions made when brought to data. The empirical findings are naturally bounded by these parametric restrictions. The method proposed in this paper is more flexible on these fronts.

Finally, the two-period rational inattention model developed in this paper is similar to the partial-equilibrium consumer problem setup in Kamdar (2019), but with only a stochastic return on capital rather than labor income. In the literature, a standard approach to solve rational inattention models is by taking a second-order approximation (e.g. Mackowiak and Wiederholt (2009), Maćkowiak *et al.* (2018), Afrouzi (2020)) and transform the problem into a Linear Quadratic Gaussian form.[6] However, such an approximation lead to symmetric and state-invariant choices of signal precision. In this paper, I solve a simple static model numerically and restrict my setup to Gaussian signals. In this setup, I show that information about the future return on capital endogenously becomes more valuable in bad states. This is because the utility loss induced by the difference between optimal saving choice under full information and that under limited information is larger in those states. This mechanism is enough to capture both the non-linearity and state dependency in agents' expectation formation process.

The rest of this paper is organized as follows: in **Section 2** I describe the empirical framework I propose and the Average Structural Function implied by such framework. In **Section 3** I introduce the method to approximate Average Structural Function using RNN and how to estimate average marginal effect of signals using the DML method. **Section 4** presents the results from applying the method to survey expectation and macroeconomic signal data. Then I propose the rational inattention model that can explain these news stylized facts in **Section 5**. And **Section 6** concludes.

2 Generic Learning Framework

In this section, I describe the empirical framework about how expectation is formed by households, which I refer to as the Generic Learning Framework. It is worth describing the similarity and key differences between this model to the standard learning models such as stationary Kalman Filter or Constant Gain Learning. In the standard models, several types of assumptions are made: (1) assumptions about information structure faced by agents that are forming expectations; (2) assumptions on identification, which involves the restrictions on unobservable error terms in the model; and (3) parametric assumptions on learning behavior. These parametric assumptions include both the underlying structure agents learn about and how learning is carried out. For example, in standard noisy information models, the perceived law of motion that the agents learn is assumed to be linear in the hidden states, and the prior and posterior beliefs on these states are structured as Gaussian. These assumptions lead to

[6] Exceptions include Sims (2006) and Flynn and Sastry (2022).

specific parametric regression methods used in different learning models. The Generic Learning Framework maintains standard assumptions on information structure and identification but imposes only minimal restrictions on the functional forms of learning. It then naturally requires the use of non-parametric or semi-parametric methods such as RNN. Such a feature also implies the Generic Learning Framework can represent a large class of learning models existing in the literature despite these models may differ in their functional forms.[7]

I introduce the Generic Learning Framework in two parts. First, I show how the agents form their expectations after observing a set of signals. This part is typically referred to as the "agent's problem". Then I describe the econometrician's information set as an observer and what she can do to learn about the agent's expectation formation process. This part is usually referred to as the "econometrician's problem".

2.1 Agent's Problem

Consider the agents observing a set of signals. These signals include both public signals that are common to each individual and private signals that are individual-specific. Denote the public signal as $X_t \in \mathbb{R}^{d_1}$ with dimension d_1 and private signal as $S_{i,t} \in \mathbb{R}^{d_2}$ with dimension d_2. An example of the public signal will be official statistics such as CPI inflation or a professional forecast of CPI inflation. An example of the private signal will be state-level inflation matched to the location agent lives or the fraction of news stories about inflation published in local newspapers.

Other than public and private signals, there is an individual level noise term denoted as $\epsilon_{i,t}$ in the agent's information set. This term represents the observational noise attached to signals in the standard noisy information model as in Woodford (2001) and Sims (2003). It can also stand for any unobserved individual-level information that is not captured by public and private signals but is used by the agent when forming expectations.

After observing the set of signals $\{X_t, s_{i,t}, \epsilon_{i,t}\}$, the agent forms expectation of variables Y_{t+1}. Denote the corresponding subjective expectation as $Y_{i,t+1|t}$.[8] The agents' expectation formation model then can be written as:

$$Y_{i,t+1|t} = \hat{\mathbb{E}}(Y_{t+1}|X_t, S_{i,t}, \epsilon_{i,t}, X_{t-1}, S_{i,t-1}, \epsilon_{i,t-1}...) = G(X_t, S_{i,t}, \epsilon_{i,t}, ...) \tag{1}$$

The formulation in (1) is a very general form of an expectation formation model. The expectation operator $\hat{\mathbb{E}}$ stands for subjective expectations formed by agents, which could be different from a statistical expectation operator. Without further assumptions, the expectations formed through this model can be non-stationary and non-tractable. To avoid these properties I make the following assumption for the Generic Learning Framework:

[7] In the Online Appendix C.1, I include an example that illustrates how this framework can represent a stationary Kalman Filter.

[8] To save notations I drop the step t, however generally speaking this could be h step expectations agents form, and it can be over any object Y.

Assumption 1. *Agents form expectations through two steps: updating and forecasting. In the updating step, agents form a finite dimensional latent variable $\Theta_{i,t}$, which follows a Stationary Markov Process:*

$$\Theta_{i,t} = H(\Theta_{i,t-1}, X_t, S_{i,t}, \epsilon_{i,t}) \tag{2}$$

In the forecasting step, they use $\Theta_{i,t}$ to form expectation:

$$Y_{i,t+1|t} = F(\Theta_{i,t}) \tag{3}$$

Where both $H(.)$ and $F(.)$ are measurable functions.

The updating step suggests that the agent holds some beliefs about the economy which can be summarized with $\Theta_{i,t}$. In each period he updates this belief from its previous level $\Theta_{i,t-1}$ with the new signals observed $\{X_t, S_{i,t}, \epsilon_{i,t}\}$. The Markov property helps to simplify the time-dependency and guarantees the tractability of the model. Stationarity makes sure the signals from history further back in time can affect expectational variables today but in a diminishing way. Furthermore, in this set up, I allow expectations to be affected by signals in the past without explicitly specifying a fixed length of memory.[9]

These two steps are commonly seen in standard learning models. For example, in stationary Kalman Filter, this is usually referred to as the "Filtering Step", where the agent uses the new signals to form a "Now-cast" variable about the current state of the economy. They will then use this "Now-cast" to form the expectation about the future using their perceived law of motion. This step is the same as the "forecasting step" in the Generic Learning Framework.

It is then worth noting that the structure of my framework described in assumption 1 covers a large class of learning models existing in the literature, other than the stationary Kalman Filter. Obviously, this formulation includes adaptive learning models where agents use only past information to form expectations.[10] It also covers models where agents get information about the future from professional forecasts through reading news stories, as in Carroll (2003). To further illustrate the flexibility of this generic framework, in Online Appendix C.1 I will take the stationary Kalman Filter that is typically used in noisy information models and a Constant Gain Learning model as two examples, and represent them in the form of the Generic Learning Framework.

In addition to Assumption 1, I also need independence assumptions on the observational noise term $\epsilon_{i,t}$. This assumption states that the noise unobservable by economists is independent of observed public and individual-specific signals as well as across individuals and time. While such an assumption is commonly made in noisy information and other learning models

[9]For example, one may want to consider a case where expectation $Y_{i,t+1|t}$ is a function of signals from a fixed window of time $\{X_t, S_{i,t}, X_{t-1}, S_{i,t-1}, ..., X_{t-h}, S_{i,t-h}\}$ Such a function is also covered by the system described by (2) and (3)

[10]See Evans and Honkapohja (2001) for example.

with unobserved noise, the economic intuition behind it is simple as well. Consider an agent wants to predict inflation, and they observe a signal on price change when they went grocery shopping. Such a signal is an imperfect measure of current inflation as it is a price change only for one or several products. Mathematically this signal can be thought of as drawn from a distribution, with the official measure of inflation being the mean of this distribution. An individual may draw the signal from the left tail or right tail of the distribution, depending on the specific product she picked up. The public signal X_t (or private signal $S_{i,t}$) is then the mean of this distribution, and $\epsilon_{i,t}$ measures the deviation of the actual signal agent observes from this mean. The assumption suggests this deviation is independent of its mean as well as across individuals and time.

Assumption 2. *The idiosyncratic noise on the public signal, $\epsilon_{i,t}$ is i.i.d across individual and time. It is orthogonal to past and future public and private signals:*

$$\epsilon_{i,t} \perp X_\tau \quad \epsilon_{i,t} \perp S_{i,\tau} \quad \forall t \leq \tau$$

$$\epsilon_{i,t} \perp \epsilon_{j,t} \quad \forall j \neq i, \quad \epsilon_{i,t} \perp \epsilon_{i,s} \quad \forall t \neq s$$

The flexible form of expectation formation in (1) together with the two assumptions summarizes the Generic Learning Framework. One can fully recover agents' expectations if $F(.)$ and $H(.)$ are known and $\{X_\tau, S_{i,\tau}, \epsilon_{i,\tau}\}_{\tau=0}^t$ and $\Theta_{i,0}$ are observable.[11]

2.2 Econometrician's Problem

Econometricians don't have all the information endowed by agents. In econometrician's problem, $\epsilon_{i,t}$ and $\Theta_{i,t}$ are typically unobservable. Furthermore, econometricians also don't have information on the functional form of $H(.)$ and $F(.)$. Denote the observable signals as $Z_{i,t} = \{X_t, S_{i,t}\}$, the econometrician only observes signals $\{Z_{i,\tau}\}_{\tau=0}^t$ and households' expectations $Y_{i,t+1|t}$.

The goal of an econometrician is to evaluate the impact of observable signals on the household's expectations. In standard learning literature, this is achieved by making structural assumptions on the expectation formation process (for example the functional forms of $F(.)$ and $H(.)$) and estimating the average marginal effect of signals or structural parameters through parametric methods. The findings from this approach are model-specific and prone to model misspecification. An alternative way to estimate the average marginal effect is by estimating the Average Structural Function (ASF) without imposing assumptions on the form of $F(.)$ and $H(.)$. Then one can use the ASF as a nuisance parameter to estimate the average marginal effect.

[11] One do not need to observe $\{\Theta_{i,\tau}\}_{\tau=1}^t$ as they can be derived from function $H(.)$, $F(.)$ and history of signals. In this sense $\Theta_{i,t}$ can be treated as part of the functional form of $H(.)$ and $F(.)$.

Average Structural Function The ASF follows from Blundell and Powell (2003). In my case the dependent variable is household expectation $Y_{i,t+1|t}$, independent variables are observed signals $\{Z_{i,\tau}\}_{\tau=0}^{t}$ and unobserved error term is $\epsilon_{i,t}$. With strict exogeneity between independent variables and unobserved errors, ASF is the counterfactual conditional expectation of dependent variable $Y_{i,t+1|t}$ given the signals $\{Z_{i,\tau}\}_{\tau=0}^{t}$. It is obtained by integrating out the unobserved i.i.d noise $\epsilon_{i,t}$:

$$
\begin{aligned}
y_{i,t+1|t} &\equiv \mathbb{E}_{\{\epsilon_{i,\tau}\}_{\tau=0}^{t}}[Y_{i,t+1|t}] \\
&= \int G(Z_{i,t}, \epsilon_{i,t}...)d\mathcal{F}_{\epsilon}(\{\epsilon_{i,\tau}\}_{\tau=0}^{t}) \\
&= \int F(H(\Theta_{i,t-1}, Z_{i,t}, \epsilon_{i,t}))d\mathcal{F}_{\epsilon}(\{\epsilon_{i,\tau}\}_{\tau=0}^{t}) \qquad (4)
\end{aligned}
$$

Where function $\mathcal{F}_{\epsilon}(.)$ is the joint CDF of all the past noise $\{\epsilon_{i,\tau}\}_{\tau=0}^{t}$. With the independence assumption 2, the ASF is equivalent to counterfactual conditional expectation function $\mathbb{E}[Y_{i,t+1|t}|\{Z_{i,\tau}\}_{\tau=0}^{t}]$.

It is immediately worth noting that the ASF can offer insight into the underlying model $G(.)$, $F(.)$ and $H(.)$ (the expectation formation process employed by agents in this case). For example, if both updating and forecasting steps follow a linear rule so that $F(.)$ and $H(.)$ are linear functions. The ASF will be linear in $Z_{i,t}$ as well. On the contrary, if the estimated ASF is highly non-linear, it suggests non-linearity in the expectation formation process.

As economists, we want to first learn features of agents' expectation formation model under the generic formulation, in this case, the structural function $G(.)$. We then want to assess how signals affect households' expectations. The ASF can be seen as a summarization of the structural functions $G(.)$, and a finite-dimensional measure of the ASF is useful to understand the properties of these structural functions. In particular, the "average derivative" of ASF can be an important measure of the marginal effects of input variables. In this paper, I define such a derivative as the average marginal effect of signals on expectations. The goal now is to estimate the ASF and the average marginal effect of the Generic Learning Framework.

3 Methodology

The estimation of Average Structural Function in forms of (4) is difficult. Under no further assumptions on updating and forecasting steps, $F(.)$ and $H(.)$ are unknown and possibly non-linear. Furthermore, the latent variable $\Theta_{i,t}$ is not directly observable, so its dimensionality is unknown.

In standard learning literature, these problems can be solved by parametric assumptions on structural function. In this paper, I take an alternative approach to directly estimate the ASF with a nonparametric method – Recurrent Neural Network. Then using the estimated ASF as a first-stage nuisance parameter, I construct a second-stage DML estimator of the average marginal effect following Chernozhukov *et al.* (2018). I start by introducing the RNN approach to estimate the Average Structural Function directly.

3.1 Estimate Average Structural Function with RNN

To estimate the ASF (4), I need a method that can capture the mapping from observed signals $\{Z_{i,t}\}$ to expectational variables flexibly. Artificial Neural Networks are known for their ability to approximate any functional forms between input and output variables.[12] However, the most popular Feedforward Neural Networks do not fit the problem well because of their inability to model time dependency between output variables and past input variables induced by the dynamic structure described before. To better fit this empirical framework, Recurrent Neural Networks are used.

RNNs are neural networks designed to model time-dependency between input and output variables. When a dynamic system describes the mapping between input and output variables, it is shown by Schäfer and Zimmermann (2006) that RNN can approximate the dynamic system of any functional form arbitrarily well. This is usually referred to as the Universal Approximation Theorem for RNN.[13] To justify that RNN can approximate the ASF of the Generic Learning Framework arbitrarily well, I need to show that the ASF (4) takes the form of a dynamic system considered by this Universal Approximation Theorem. Theorem 1 shows that the ASF can be well-approximated by a dynamic system of equations with a finite-dimensional $\theta_{i,t}$.

Theorem 1. *For any dynamic system described in* (2) *and* (3)*, with assumptions 1 and 2 hold, input vector $Z_{i,t} \in \mathbb{R}^s$, where $s = d_1 + d_2$, and output vector $Y_{i,t+1|t} \in \mathbb{R}^l$. Denote the average structural function* (4) *as:*

$$y_{i,t+1|t} \equiv g(\{Z_{i,\tau}\}_{\tau=0}^{t}, \theta_{i,-1}) \tag{5}$$

There exists a finite dimensional $\theta_{i,t} \in \mathbb{R}^d$, a continuous function $f : \mathbb{R}^d \to \mathbb{R}^l$ and a measurable function $h : \mathbb{R}^s \times \mathbb{R}^d \to \mathbb{R}^d$ s.t. the average structural function described in (4) *can be written as a dynamic system:*

$$\begin{aligned} y_{i,t+1|t} &= f(\theta_{i,t}) \\ \theta_{i,t} &= h(\theta_{i,t-1}, Z_{i,t}) \end{aligned} \tag{6}$$

Notice equation (5) is an alternative representation of ASF (4). In (5) the inputs of function $g(.)$ are the history of observed signals $\{Z_{i,\tau}\}_{\tau=0}^{t}$ and the initial levels of θ at time

[12] See the Universal Approximation Theorem addressed in Hornik *et al.* (1989).

[13] According to the Universal Functional Approximation Theorem (See Hornik *et al.* (1989) for the results for Feed Forward Networks and Schäfer and Zimmermann (2006) for Recurrent Networks), a single layer neural network with sigmoid activation function can approximate any continuous function. The result is extended to nerual networks with Rectifier Linear (ReLu) activation function by Sonoda and Murata (2015).

$t = 0$, $\theta_{i,-1}$. The unobserved noise $\epsilon_{i,t}$ are integrated out and the information contained in hidden states $\Theta_{i,t}$ is captured by the construction of $\theta_{i,t}$. The proof of Theorem 1 can be found in Appendix A. I then use a state-of-art RNN with Rectifier Linear (ReLu) activation function to approximate the ASF (4) derived from the Generic Learning Framework.[14] Now denote the class of functions in RNN $\mathcal{G}_{foh}^{RNN}$, the estimator is computed by minimizing the sample mean squared errors:

$$\hat{g}_{rnn} := \arg\min_{g_w \in \mathcal{G}_{foh}^{RNN}} \sum_{i,t} \frac{1}{2}\Big(Y_{i,t+1|t} - g_w(\{Z_{i,\tau}\}_{\tau=0}^{t}, \theta_{i,-1})\Big)^2$$

In Theorem 1 the alternative representation (5) also shows with the same realization of $Z_{i,t}$, $y_{i,t+1|t}$ may differ at different point of time. Moreover, such a difference comes from the accumulation of signals they see, $\{Z_{i,\tau}\}_{\tau=0}^{t}$ rather than the underlying structural functional forms $f(.)$ and $h(.)$. In other words, such a flexible formulation allows for endogenous time-varying marginal effect of signals $Z_{i,t}$. This point will become more clear when I introduce the average marginal effect.

3.2 Estimate Average Marginal Effect with DML

Now I turn to the other object of interest: the average marginal effect of a particular signal. This is the mean of gradient for Average Structural Function $g(\{Z_{i,\tau}\}_{\tau=0}^{t}, \theta_{i,-1})$:

$$\beta = \mathbb{E}[\nabla g(\{Z_{i,\tau}\}_{\tau=0}^{t}, \theta_{i,-1})] \tag{7}$$

Or for a single signal $z_{j,i,t}$ which is the j-th element in vector $Z_{i,t}$, this can be written as:

$$\beta^j = \mathbb{E}[\frac{\partial g(\{Z_{i,\tau}\}_{\tau=0}^{t}, \theta_{i,-1})}{\partial z_{j,i,t}}] \tag{8}$$

The equation (7) can be thought of as a moment condition used to estimate β. With the functional estimator obtained from RNN, a plug-in estimator of β is available by computing the sample mean of the partial derivative using estimator of conditional expectation function: $\mathbb{E}_n[\nabla \hat{g}_{rnn}(\{Z_{i,\tau}\}_{\tau=0}^{t}, \theta_{i,-1})]$. However, such an estimator typically has two problems: (1) when regularization is used in RNN, which is the case here, the estimate using moment condition (7) is usually biased; (2) the functional estimates obtained by Machine Learning (RNN in this case) methods typically have slower than $\sqrt{n}$ convergence speed. This makes the estimate not well-behaved asymptotically, thus making inference hard.[15]

One way to solve these problems is to use the DML method as proposed by Chernozhukov *et al.* (2018) and Chernozhukov *et al.* (2017). I can form the estimation problem as a semi-parametric moment condition model with a finite-dimensional parameter of interest,

[14] The RNN approximate dynamic systems (6) by constructing representations of $\theta_{i,t}$ as well as $f(.)$ and $h(.)$.

[15] These issues are well discussed in Chernozhukov *et al.* (2018), they also propose ways to solve these problems. One way they proposed is the DML approach, which is what I follow to estimate the average marginal effect in this paper.

β; infinite-dimensional nuisance parameter η (including functional estimator from Machine Learning methods, $\hat{g}_{rnn}$ in this case), and a known moment condition $\mathbb{E}[\psi(W;\beta,\eta)]$. The benefits of this approach are two folds, it first corrects for biases in the estimator, and it also offers a way to obtain valid inference on the estimator. The plug-in estimator is usually biased and not asymptotically normal because the construction of the estimator of β involves the regularized nuisance parameters obtained by Machine Learning methods (in this case RNN). This Machine Learning estimator usually has a convergence speed slower than $\sqrt{n}$ and makes the estimator on β exploding as sample size goes to infinity. Using orthogonalized moment conditions solves this problem because the moment conditions used to identify β are locally insensitive to the value of the nuisance parameter. This allows me to plug in noisy estimates of these parameters obtained from RNN.

The estimator $\hat{\beta}$ is then $\sqrt{n}$ asymptotic normal under appropriate assumptions on estimate of nuisance parameter $\hat{\eta}$ and the moment condition. These conditions typically require the moment condition to be (Near) Neyman Orthogonal; function $\psi(.)$ to be linearizable and a fast enough convergence speed of nuisance parameter.[16]

The convergence speed requirement for Neural Networks with ReLu activation functions is verified in Farrell *et al.* (2021). Then following the concentrating-out approach in Chernozhukov *et al.* (2018), I can derive the Neyman Orthogonal Moment Condition for β^j:

$$\mathbb{E}[\beta^j - \frac{\partial g(\{Z_{i,\tau}\}_{\tau=0}^t, \theta_{i,-1})}{\partial z_{j,i,t}} + \frac{\partial ln(f_z(\{Z_{i,\tau}\}_{\tau=0}^t, \theta_{i,-1}))}{\partial z_{j,i,t}}(Y_{i,t+1|t} - g(\{Z_{i,\tau}\}_{\tau=0}^t, \theta_{i,-1}))] = 0 \quad (9)$$

The nuisance parameters associated with moment condition (9) then include both the average structural function $g(.)$ as well as the joint density function $f_z(\{Z_{i,\tau}\}_{\tau=0}^t, \theta_{i,-1})$. One complication here is the joint density function could be high-dimension, and it includes both current and past signals. Here I make an extra assumption that the signal Z follows a VAR(1) so that to get the estimate of the partial derivative of log density, I only need to estimate the joint density of $f_z(Z_{i,t}, Z_{i,t-1})$. The joint density is then obtained using higher-order multivariate Gaussian Kernel Density Estimation with bandwidth chosen according to Silverman (1986) to guarantee the appropriate convergence speed of the density estimator. The estimator of β^j is obtained by the following steps:

1. Estimate nuisance parameter $\eta = \{g, f_z\}$. g is estimated by RNN and f_z is estimated by Gaussian Kernel Density Estimation. Denote the estimates as $\hat{g}_{rnn}$ and $\hat{f}_z$ respectively.

2. Obtain estimate of average structural function from computing derivative numerically:
$$\frac{\partial \hat{g}_{rnn}}{\partial z_{j,i,t}} = \lim_{\delta \to 0} \frac{\hat{g}_{rnn}(Z_{i,t} + \Delta_j/2, \{Z_{i,\tau}\}_{\tau=0}^{t-1}, \theta_{i,-1}) - \hat{g}_{rnn}(Z_{i,t} - \Delta_j/2, \{Z_{i,\tau}\}_{\tau=0}^{t-1}, \theta_{i,-1})}{\delta} \quad (10)$$
Where $\Delta_j \in \mathbb{R}^s$ is a vector of zeros, with jth element being δ.

[16]For the formal formulation of semi-parametric moment condition model, derivation of Neyman Orthogonality condtion and convergence speed requirements of nuisance parameter, refer to Appendix B

3. The estimate of $\frac{\partial ln(\hat{f}_z(Z_{i,t},Z_{i,t-1}))}{\partial z_{j,i,t}}$ is obtained similarly using numerical derivatives.

$$\frac{\partial ln(\hat{f}_z(\{Z_{i,\tau}\}_{\tau=0}^{t},\theta_{i,-1}))}{\partial z_{j,i,t}} = \lim_{\delta\to 0}\frac{\hat{f}_z(Z_{i,t}+\Delta_j/2,Z_{i,t-1})-\hat{f}_z(Z_{i,t}-\Delta_j/2,Z_{i,t-1})}{\delta\hat{f}_z(Z_{i,t},Z_{i,t-1})} \tag{11}$$

4. Then the DML estimate is given by:

$$\hat{\beta}^j = \frac{1}{N}\sum_i\frac{1}{T}\sum_t\underbrace{[\frac{\partial\hat{g}_{rnn}(\{Z_{i,\tau}\}_{\tau=0}^{t},\theta_{i,-1})}{\partial z_{j,i,t}} - \frac{\partial ln(\hat{f}_z(\{Z_{i,\tau}\}_{\tau=0}^{t},\theta_{i,-1}))}{\partial z_{j,i,t}}(Y_{i,t+1|t}-\hat{g}_{rnn}(\{Z_{i,\tau}\}_{\tau=0}^{t},\theta_{i,-1}))]}_{\equiv\hat{\beta}^j_{i,t}} \tag{12}$$

4 Application to Survey Data

In this section, I use survey data of expectation and a rich set of macroeconomic signals to estimate the Average Structural Function of the Generic Learning Framework. There is growing literature using survey data to estimate learning models. The respondents in the surveys that researchers usually differ. The most widely explored expectations are those from households and professionals. In this paper, I focus on households' expectations from the US, and I use professional forecasts (SPF) as a signal that households can utilize to form their expectations, similar to the idea of Carroll (2003). However, in my empirical method, I allow households' expectations to respond to information in SPF in a flexible way.

4.1 Data Description

Table 1 summarizes the data on expectations and signals used to estimate the generic learning model as well as the notations being used.

For outcome variable $Y_{i,t+1|t}$ I use Reuters/Michigan Survey of Consumers (MSC). It is a monthly survey for a representative sample of US households with a preliminary interview usually conducted at the beginning of the month. The survey asks about the respondent's one-year-ahead expectations on various macroeconomic aspects. In this paper, I include four expectational variables of interest: (1) expected inflation rate, denoted as $\hat{\pi}_{i,t+1|t}$; (2) whether the economic condition will be better, denoted as $\Delta\hat{y}_{i,t+1|t}$; (3) whether unemployment rate will increase, denoted as $\Delta\hat{u}_{t+1|t}$; (4) whether the interest rate will increase $\Delta\hat{r}_{t+1|t}$.

I include two sets of public signals X_t. One is the realized economic statistics from the Federal Reserve of St. Louis. These signals contain information about the current state of the economy. Another set of public signals I consider is the professional forecasts from the Federal Reserve of Philadelphia. These signals are considered as containing information about the future because they usually lead and Granger-Cause the predicted macroeconomic variables.[17]

[17]See Carroll (2003) for details

Table 1: Data Description: some key notations

Input variable $(X_t, S_{i,t})$	Variable and Notation	Source
Macro variable	CPI: π_t, unemployment: Δu_t, Federal Funds Rate: r_t, real GDP growth: $\Delta rgdp_t$, Real Oil price: o_t Stock price index: $stock_t$	FRED
Professional Forecasts	CPI: $F_t \pi_{t+1}$, unemployment change: $F_t \Delta u_{t+1}$, short term Tbill: $F_t \Delta r_{t+1}$, real GDP growth: $F_t \Delta rgdp_{t+1}$ anxious index: $F_t rec_{t+j}$	Survey of Professional Forecasters (Philadelphia FED)
Individual Signals	regional CPI: $\pi_{i,t}$, regional unemployment: $\Delta u_{i,t}$ news on recession: $Nrec_{i,t}$ news on inflation: $N\pi_{i,t}$ news on boom: $Nboom_{i,t}$ news on interest rate: $Nr_{i,t}$	Bureau of Labor Statistics, LexisNexis Uni
Individual Lag Expectation	inflation rate: $\hat{\pi}_{i,t\|t-1}$ change of economic condition: $\Delta \hat{y}_{i,t\|t-1}$ unemployment change: $\Delta \hat{u}_{i,t\|t-1}$ interest rate change: $\Delta \hat{r}_{i,t\|t-1}$	Michigan Survey of Consumers
Output variable $(Y_{i,t+1\|t})$	**Variable and Notation**	**Source**
Expectational Variable	inflation rate: $\hat{\pi}_{i,t+1\|t}$ change of economic condition: $\Delta \hat{y}_{i,t+1\|t}$ unemployment change: $\Delta \hat{u}_{i,t+1\|t}$ interest rate change: $\Delta \hat{r}_{t+1\|t}$	Michigan Survey of Consumers

Then in individual-level signals $S_{i,t}$, I include the local unemployment rate and CPI inflation matched with the individual in MSC according to their location information. I also include the intensity of news story reports on recessions, inflation and interest rates at both local and national level.[18] The idea that information about future flows from professional forecasts to households through media reports can be dated back to Carroll (2003) and has lots of follow-up researches.[19] I include the news measure as RNN allows for interaction between input variables, so the transmission of information can also be captured. I also include the lagged expectations of households as extra inputs. The assumption that observational noise is uncorrelated across time guarantees the lagged expectation won't be correlated with the unobserved error term $\epsilon_{i,t}$.

Because the panel component of MSC only has two waves for each individual, whereas capturing the latent state accumulated by observing the history of signals requires a longer time dimension. For this reason, the data set is compiled as a synthetic panel. Each synthetic agent is grouped by its social-economic status, including income quantile, region of living, age, and education level. Because these four characteristics were found significantly affect expectation by Das *et al.* (2019). The baseline sample I am using is quarterly from 1988 quarter 1 to 2019 quarter 1. The length of the sample is due to the availability of data on news stories.[20] The frequency of data is quarterly because professional forecasts are quarterly data.

4.2 Results

Estimation of functions with RNN usually requires selection of network architecture. Because of the superior performance in applications of modern neural networks, I choose Rectified Linear (ReLu) Activation functions for all the layers in RNN and use Long-Short Term Memory (LSTM) recurrent layer. It is worth noting the requirements for convergence speed offered by Farrell *et al.* (2021) are also for neural networks with ReLu activation functions, and the width (number of neurons) and depth in my baseline architecture of RNN satisfy these requirements. The rest configurations of hyper parameters are chosen using a standard K-Fold Cross Validation, in my case $K = 6$.[21] Table 2 summarizes the architecture of RNN I use.

[18] I scraped volume of reports on related macroeconomic topics from TV news scripts and local newspaper articles. Following PFAJFAR and SANTORO (2013) I construct a measure of news coverage on these topics by computing the number of news stories on each topic (for example, news about inflation) in each quarter as a fraction of total news stories in the same quarter, and I include only news with more than 120 words to exclude short reviews or notice. The data is available from LexisNexis Database.

[19] See PFAJFAR and SANTORO (2013) and LAMLA and MAAG (2012) for examples.

[20] Prior to 1988, there are too few local published newspapers included in LexisNexis Database.

[21] I also tried RNN with smaller width and no regularization (dropout) as well as more complex architectures, the results don't change qualitatively. To assess the stability of the neural networks I also tried with multiple random initial weights and the results are stable across different initial weights used.

Table 2: Architecture RNN

Tuned Hyper Parameter	Configuration
Num. of Recurrent Neurons	32
Feed-forward Neurons	20
Dropout on recurrent layer	0.5
Epochs	200
Learning Rate	$1e^{-6}$
Depth	2(4)
Un-tuned Hyper Parameter	Configuration
Type of Recurrent Layer	Long-Short Term Memory (LSTM)
Activation Function:	ReLu

* Tuned hyper parameters are picked using 6-Fold cross-validation across individuals. There is 1 layer of recurrent neurons that are connected to 1 layer of feed-forward neurons. Because each one LSTM layer contains 3 layers of neurons, this makes the actual depth of network being 4. It is worth noting such depth satisfies the requirement for fast enough convergence of estimated Average Structural Function so that functional estimators from this Neural Network can be used to obtain inference on DML estimators.

It is important to note the estimated ASF has a 4-dimensional output, and more than 20 inputs are considered. The ASF and marginal effects can be presented in each signal-expectation pair. In this paper, I will only focus on the impact of signals on expectations regarding the same subjects, which I refer to as "self-response". For example, I will look at the impact of the realized unemployment rate on unemployment expectations for the future.[22]

The estimation procedure described in **Section 3** involves several steps. In this subsection, I present results progressively following those steps. I first show the estimated ASF from the baseline RNN described in Table 2. Then I present the time-varying marginal effects of macroeconomic signals implied by the estimated ASF. I interpret this finding as an "attention-shift" of households from signals about the past and current state of the economy to signals that contain information about the future. Then I obtain the DML estimator of marginal effects with inference and perform tests to show that such an "attention shift" is statistically significant. Finally, I explore reasons for the "attention shift" by doing a decomposition of the time-varying marginal effects of interest. The identified key driving forces are then used in the rational inattention model I proposed to rationalize findings from RNN.

[22]Another interesting direction is to examine "cross-response", for example, how signals on inflation affect unemployment expectation. This direction is explored in a somewhat related work Hou (2020).

4.2.1 Estimated Average Structural Function

For an easy representation of ASF in (4), denote the signal considered in the input dimension as x_t, and the one-dimensional output is the expectational variable on the same subject, denoted as $E_t x_{i,t+1}$. Then use $Z_{i,t}^{-x}$ to represent contemporaneous signals other than x_t. Following from (6), the estimated functional estimator can be expressed as the following function:

$$E_t x_{i,t+1} = \hat{g}_x(\theta_{i,t-1}, Z_{i,t}^{-x}, x_t) \tag{13}$$

Now take unemployment as an example subject. Figure 1 plots the average structural function of expected probability for future unemployment rate increase, along the signal on change of actual unemployment rate. Following (13), this function can be written as:

$$E_t \Delta u_{t+1} = \hat{g}_u(\theta_{i,t-1}, Z_{i,t}^{-u}, \Delta u_t) \tag{14}$$

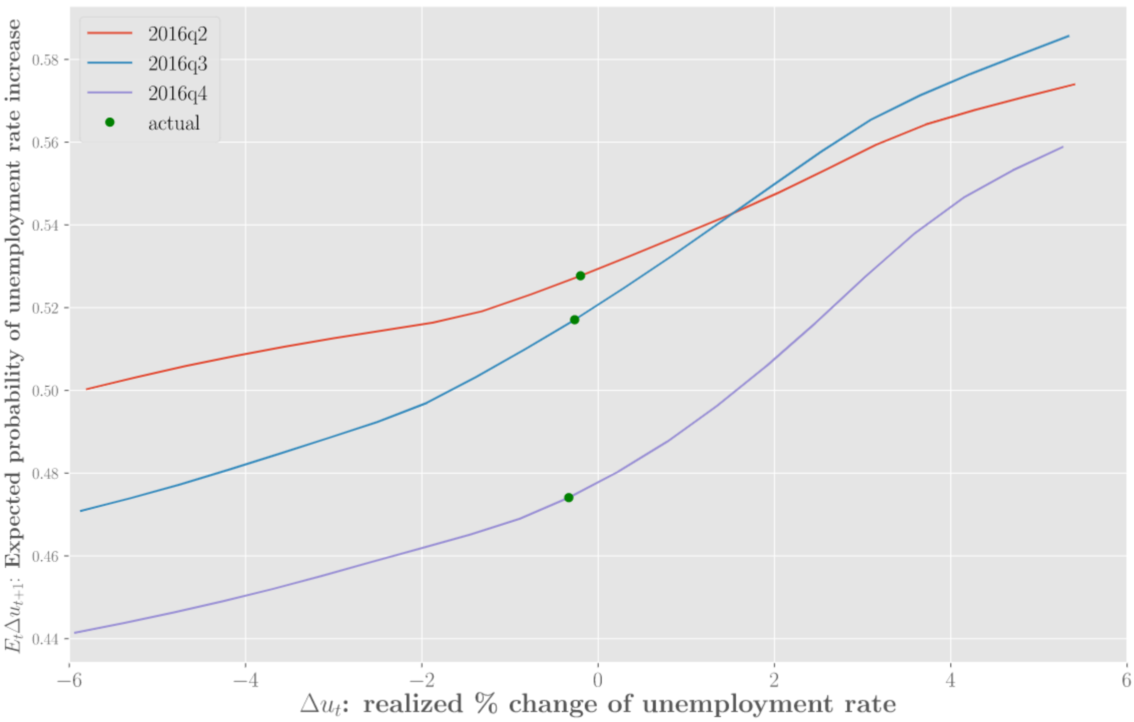


Figure 1: Average of expected probability for unemployment rate increase $E_t \Delta u_{t+1}$ as function of realized unemployment rate change Δu_t, at different point of time. Purple curve: 2016q4, blue curve: 2016q3, red curve: 2016q2. The dot on each curve represents the prediction from estimated function when actual data in that period is input.

In Figure 1, the function (14) is plotted at three different points of time: quarters 2,3, and 4 in 2016. This graph shows that at different points of time, households may form different expectations in response to the same signal on the realized unemployment rate change. However, such a difference comes from either the hidden states ($\theta_{i,t-1}$) they accumulated from observing a different path of signals or the interactions between newly observed signals $Z_{i,t}$.[23] In other words, any state dependency I find with the estimated ASF is a result of the signals households observed. This is a crucial implication of the model that comes from the flexibility of the Generic Learning Framework and RNN method.

From Figure 1 we see the estimated ASF at different points of time are highly non-linear. When the unemployment rate falls the curve is flat and expectations of the unemployment rate respond only mildly. Whereas when unemployment rates increase the curve becomes steeper and then flat again when the change of unemployment rate is really high. As a result of this non-linear response, the ASF appears to be asymmetric. Take 2016 quarter four as an example. The curve implies that if unemployment had increased by 1.6% instead of falling by 0.4% (a "bad news"), the ASF predicts households will be 5% more likely to believe unemployment will increase in the future. However, if unemployment decreased further by 2.4% (a "good news"), households will only be 3% less likely to expect the unemployment rate to go up.[24]

To assess the significance of the asymmetry from the estimated ASF above, I turn to estimate average deviations of expectation and obtain valid inference using DML as described in Appendix B.1.

$$\gamma_\delta = \mathbb{E}[g(Z_{i,t} + \delta, \{Z_{i,\tau}\}_{\tau=0}^{t-1}, \theta_{i,-1}) - g(Z_{i,t}, \{Z_{i,\tau}\}_{\tau=0}^{t-1}, \theta_{i,-1})] \tag{15}$$

The average deviation is defined in equation (15), it describes the average (across $\{Z_{i,\tau}\}_{\tau=0}^{t}$) change of expectational variable when signal $Z_{i,t}$ increase by δ, relative to its original level. As this needs to be done for each output-input pair, I again focus on the pairs in which the output expectational variable and input signal variable are on the same subject (the "self-response").

In Figure 2 I plot the average deviation for all four expectational variables along with the corresponding signals. In each case, I consider 20 different values of δ symmetrically centered around 0. For each point estimate at δ, I present the 95% confidence interval. Panel (a) shows the average deviation for unemployment expectation along with the change in unemployment signal. It shows similar patterns as in the estimated ASF presented in Figure 1: the expectations are more responsive to unemployment rate surge and the responses are more muted when the unemployment rate falls or becomes too high. The confidence interval shows the asymmetry is significant.

Comparing all four panels in Figure 2, I find such a non-linearity shows up consistently in cases of unemployment expectation and economic condition expectation. In panel (b) when Δy falls drastically, the slope of ASF becomes flat, the same as the case when the unemployment

[23] Given that they are close to each other in time (should have similar hidden state accumulated) and current Δu_t is roughly at the same level. The primary reason for the level difference here is that the lag expectation $E_{t-1}\Delta u_t$ was higher in 2016q2 and q3. The fact that expected unemployment is gradually falling illustrates how expectation is slowly adjusting downwards when the actual unemployment rate keeps falling($\Delta u_t < 0$) throughout the three quarters plotted.

[24] Such a pattern will not be seen in a linear model if the underlying expectation formation model is linear in signals, the ASF will be linear as well.

signal is high in panel (a). Then it gets steeper as δ becomes closer to 0, and gets flat again when δ keeps increasing and becomes positive. On the other hand, in panels (c) and (d), which correspond to inflation and interest rate expectation as functions of inflation and interest rate signal, the relationships are closer to linear.

These observations lead to two major patterns among all the findings in my application of RNN to survey data: (1) findings are most stark in cases with expectations on economic condition (e.g. unemployment change $E_t\Delta u_{i,t+1}$ and economic condition change $E_t\Delta y_{i,t+1}$), and these results are consistent between these two measures. One can think of unemployment (expectation or signal) as a negative counterpart of economic condition/RGDP. (2) findings on expected inflation and interest rate are more consistent with those from existing literature. These patterns also hold for my later findings on time-varying and average marginal effects. For these reasons, I will focus on presenting results with the expected economic condition, $E_t\Delta y_{i,t+1}$, from now on. [25]

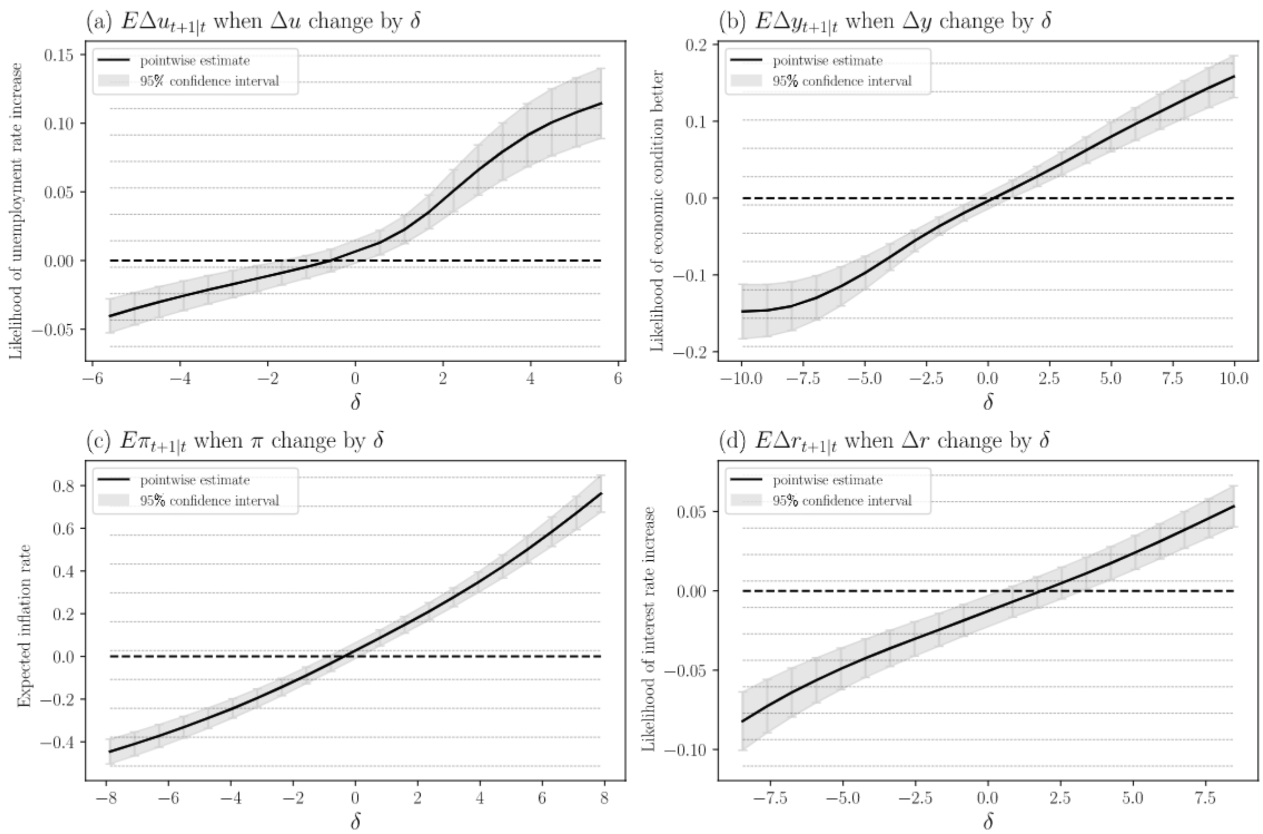


Figure 2: Average deviation of four expectational variables in response to signals on themselves. Panel (a): expected likelihood of unemployment increase as unemployment signal change by δ. Panel (b): expected likelihood of economic condition be better as real GDP signal change by δ. Panel (c): expected inflation rate as inflation signal change by δ. Panel (d): expected likelihood of interest rate increase as interest rate signal change by δ.

[25] For results on the other three expectational variables, I include the results in Online Appendix D.1.

4.2.2 State-dependent Marginal Effect

Following from the estimated ASF in (13), I can define the average (across individual) time-specific marginal effect of signal x on expectational variable Ex as:

$$\beta_{x,t}^{Ex} = \mathbb{E}_n[\frac{\partial \hat{g}_x(\theta_{i,t-1}, Z_{i,t}^{-x}, x_t)}{\partial x_t}] \tag{16}$$

This marginal effect is different at each point of time t for the same reason as discussed before: different internal state $\theta_{i,t-1}$ and contemporaneous signal $Z_{i,t}$. It describes on average how responsive the expectation $E_t x_{t+1}$ is to the change of signal x_t at time t after observing all signals up to that time. It can then be interpreted as weights applied to signals following the standard learning literature. In the rest of this paper, I will use weights and marginal effects interchangeably. If the underlying learning model doesn't feature endogenous states or interactions between signals and states, for example, stationary Kalman Filter, this marginal effect will not have a time-varying slope.[26] In this section, I show profound time-variation in the average marginal effect of signals on expectations about the economic condition. Specifically, such a time variation implies households' attention to signals is cyclical: they put lower weights on signals about current and past states and, at the same time, more weight on signals about the future during periods with bad economic conditions.

Before I proceed to these results, it is useful to define two related notions: (1) signal about the past and signal about the future; (2) bad times and ordinary times.

Signals about past v.s. future: Agents can acquire information about the current state of the economy from macroeconomic statistics. They get this information either directly as it is publicly available or partially through daily activities. I will use realized key macroeconomic variables as a proxy for the signal about the past. Expectations formed majorly relying on this information are then treated as adaptive. For signals about the future, I follow Carroll (2003) and use consensus (average) expectation from the Survey of Professional Forecasters as a proxy. Information about the future can take the form of news or anticipated shocks as in Beaudry and Portier (2006) and Barsky and Sims (2012), and it flows into the household's information set through news media as suggested in Carroll (2003).

Bad time v.s. ordinary time: For periods characterized as "bad time", I take the ones that have at least 2 consecutive quarters with the unemployment rate increasing: 1990q3-1992q3, 2001q1-2002q4 and 2007q3-2010q3.[27] The results will not change qualitatively if I use

[26] It is closely related to the curvature of estimated ASF presented in the previous section but not related to the level difference. For example, in the stationary Kalman Filter, its ASF recovered by RNN may still be different in levels at each point of time.

[27] Notice the unemployment rate change I use, Δu_t is year-to-year unemployment rate change. I pick the quarters that have $\Delta u_t > 0$ with 2 consecutive quarters around it also have $\Delta u_t > 0$. This choice is because I

the NBER recession dates to measure "bad time".[28]

I then present the time-specific marginal effect from (16) of signals on real GDP growth. I consider both signals about the past and future. In Figure 3, the color bars in the top panel are the marginal effects of real GDP growth signal, $x_t = \Delta y_t$, on expected economic condition next year; those in the bottom panel are the marginal effects of professionals' forecasts about real GDP growth next year, $x_t = F_t \Delta y_{t+1}$, on expected economic condition. Both marginal effects are normalized by standard deviations for ease of comparison.

The color bars in each panel stand for the corresponding marginal effect at that point in time. A red color means a positive marginal effect; a blue color means a negative marginal effect, and white means the marginal effect is zero. The color map is on the right side of each panel, and the scale stands for normalized marginal effect. For example, 0.1 on the color map means when signal x_t changes by 1 standard deviation, the corresponding expectation changes by 0.1 standard deviations. This is then represented by a dark red color bar in the graph. The darker the color, the bigger the magnitude of the marginal effect. The solid black line is the series of signal x_t at which I evaluate the marginal effect. The dotted area is the NBER recession episode.

In general, both higher real GDP growth and higher forecasted growth by professionals make households predict better economic conditions. The maximum of marginal effect of real GDP growth is 0.24 in 1996 quarter 1, which indicates 1 standard deviation increase of real GDP growth (approximately 1.66%) leads to a 0.24 standard deviation increase in expected business condition (on average 0.125 more likely to believe the economic condition to be better).

One key observation comes from comparing the top panel to the bottom. In panel (a), the pale color during recession periods in panel (a) suggests that the marginal effect of the past signal is close to zero or negative. In contrast, the red color bars indicate the marginal effects are usually sizeable during non-recession episodes. On the other hand, in panel (b), the patterns for marginal effects on the future signals are the opposite: higher during the recession period than in ordinary periods. Such an observation indicates that households are more sensitive to signals about the past during ordinary periods and put more weight on signals about the future when the economic condition gets worse. It is also important to note that it does not necessarily mean they are more pessimistic during bad times because negative or close-to-zero marginal effects do not mean worse expectations of economic conditions, rather

use a year-to-year change in unemployment rate as the measure of unemployment rate signal, and this measure appears to return to zero 2 to 4 quarters after the day that marks the end of NBER recessions. Using such a characterization shows weights on signal change are related to the signal itself rather than an external definition of "bad period" as it is reasonable to think that households won't have the information on the end date of NBER recessions when they form expectations around the same time. The announcement typically comes out at least 2 quarters after the official end day of the NBER recession.

[28] These results using NBER recession dates as robustness check is included in Online Appendix D.2.

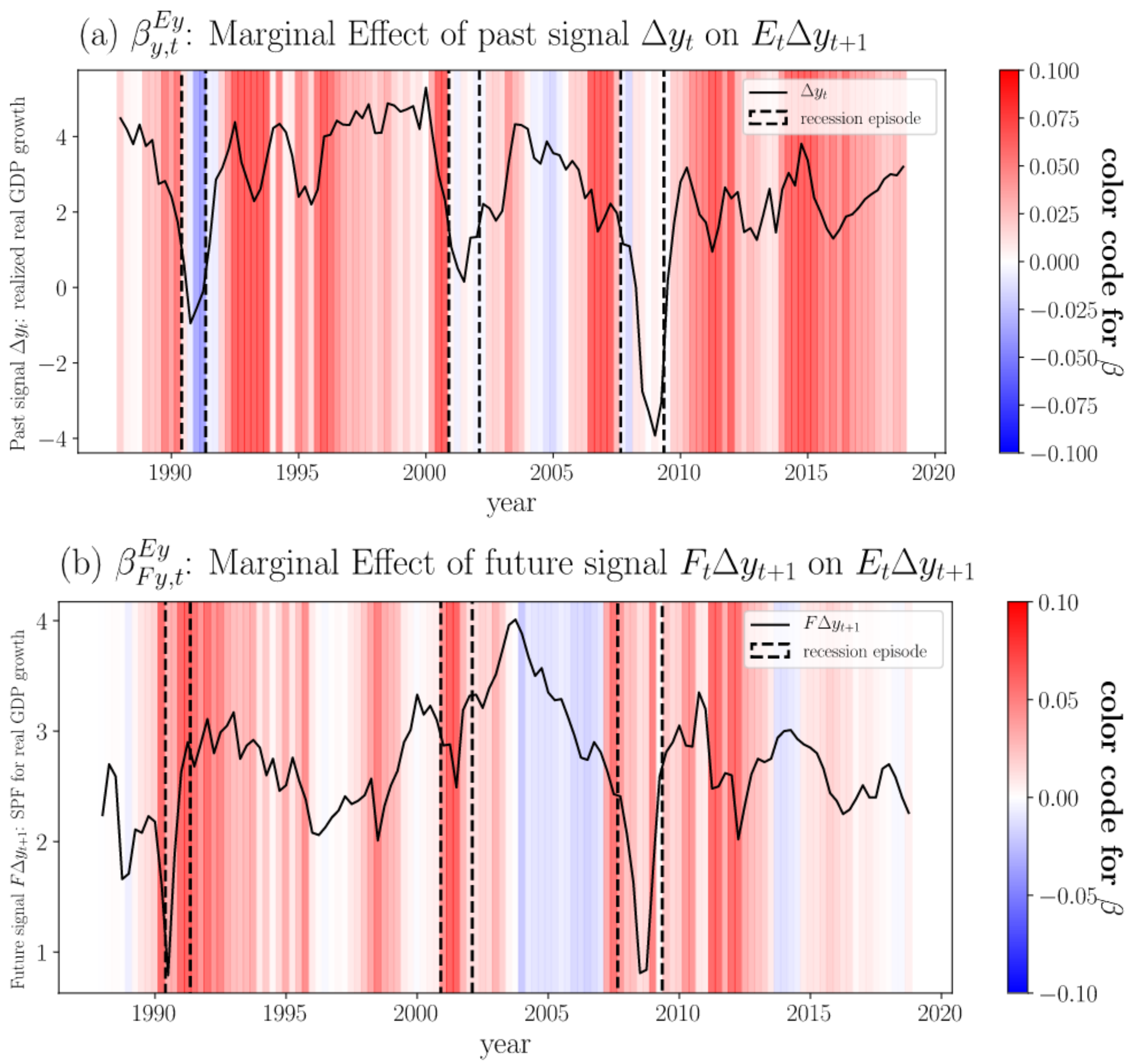


Figure 3: Color bars in panel (a): the marginal effects of real GDP growth signal Δy_t on expected economic condition next year $E\Delta y_{t+1|t}$. Panel (b): the marginal effects of professionals' forecasts about real GDP growth next year $F\Delta y_{t+1|t}$ on expected economic condition. Red color: positive marginal effect; blue color: negative marginal effect. Black solid line: data on the signal considered.

it means the expectation is less responsive to the signal considered.

Such a finding is obviously at odds with models that impose time and state invariant weights on different signals, such as constant gain learning and model with stationary Kalman Filters. It is more consistent with the case that agents shift their attention to signals about the future thus becoming more "forward-looking" during bad times in the economy. Moreover, such a finding does not only exist in expectation and signals on economic condition Δy, but it also qualitatively holds for expectation and signals on unemployment status Δu. In the next section, I follow Chernozhukov *et al.* (2018) and obtain the DML Estimator on average marginal effects (AME) in bad and ordinary times. The DML method helps to correct the potential biases and allows me to assess whether the AMEs are different in bad and ordinary times.

4.2.3 DML Estimator of Average Marginal Effects

I compute the DML Estimator following the procedures described in Section 3.2. Table 3 reports the estimated AME of past and future signals on expected economic conditions and expected unemployment rate change. I separate the time-varying marginal effects into two groups, β_{rec} denotes the average marginal effect during "bad periods" defined before. And β_{ord} denotes the average marginal effect in periods other than the bad episodes. I then perform a Wald test on $\beta_{rec} = \beta_{ord}$, the p-value is also reported in the table.

Table 3: Average Marginal Effect of Past and Future Signals on Expectation

Expectation:		$E\Delta y_{t+1\|t}$			$E\Delta u_{t+1\|t}$		
	Signal	β_{bad} (std)	β_{ord} (std)	$\beta_{bad} = \beta_{ord}$ (p-val)	β_{bad} (std)	β_{ord} (std)	$\beta_{rec} = \beta_{ord}$ (p-val)
Future Signal	$F_t\Delta u_{t+1}$	**−0.037*** **	0.009**	< 0.01	**0.029*** **	0.007***	< 0.01
		(0.004)	(0.002)		(0.003)	(0.002)	
	$F_t\Delta y_{t+1}$	**0.049*** **	0.016***	< 0.01	**−0.022*** **	−0.009***	< 0.01
		(0.005)	(0.003)		(0.002)	(0.001)	
	$F_t\Delta r_{t+1}$	0.026***	0.025***	0.92	−0.022***	−0.021***	0.79
		(0.007)	(0.004)		(0.004)	(0.002)	
	$F_t\pi_{t+1}$	0.014***	0.003**	< 0.01	−0.008***	0.000	< 0.01
		(0.002)	(0.001)		(0.002)	(0.001)	
Past Signal	Δu_t	−0.006	**−0.021*** **	0.04	0.005	**0.012*** **	0.08
		(0.006)	(0.004)		(0.004)	(0.002)	
	Δy_t	0.004*	**0.017*** **	< 0.01	−0.006***	**−0.01*** **	0.04
		(0.003)	(0.001)		(0.001)	(0.002)	
	Δr_t	0.002	0.003***	0.80	0.004*	0.004**	0.99
		(0.002)	(0.001)		(0.002)	(0.001)	
	π_t	−0.007***	−0.008***	0.67	−0.000	0.001	0.40
		(0.003)	(0.002)		(0.001)	(0.001)	

* ***,**,*: Significance at 1%,5% and 10% level. β_{bad} is average marginal effect in bad periods defined before, β_{ord} is average marginal effect in ordinary period. $\beta_{bad} = \beta_{ord}$ is test on equality between average marginal effects, its p-value is reported for each expectation-signal pair. Bold estimates denote the marginal effect with significantly bigger magnitude. Standard errors are adjusted for heteroskesticity and clustered within time.

The key message from Table 3 can be seen by comparing the marginal effects of the same signal between bad and ordinary periods. For future signals on unemployment and real GDP growth, their marginal effects always have a bigger magnitude during bad episodes, whereas the effects of past signals are always bigger in ordinary episodes. The p-values on the Wald test with the null hypothesis: $H_0 : \beta_{bad} = \beta_{ord}$ range from 0.08 to less than 0.01 for these signals, which suggests the difference of marginal effects is statistically significant at least at 10% level. However, the same pattern does not hold true for signals on inflation and interest rate, with the exception of the future signal on inflation. In fact, average marginal effects on these signals are either insignificant or with small magnitudes. These results show that the attention shift documented before is statistically significant and it only exists for expectations and signals on real economic activities. In other words, they are more adaptive learners when economic conditions are stable and become more forward-looking when the situation gets worse.

4.2.4 Decomposing Time-varying Marginal Effect

Now I have shown that households put more weight on signals from professional forecasters in bad times; meanwhile, they rely less on realized macroeconomic statistics. However, the explanation for such a weight shift remains unclear. As the time variation is only created by inputs to the RNN, I can use the trained ASF to decompose the contributions coming from different sets of input signals. I separate input signals for RNN into four categories: signals about economic conditions, signals about inflation, signals about the interest rate, and measures of news exposure about economic conditions.

As estimated ASF is non-linear, a proper way for variance decomposition is to use the Law of Total Variance following Isakin and Ngo (2020). I compute the direct contribution to the time-varying marginal effects of past and future signals on expectations related to economic conditions (those regarding Δu and Δy) for each of the four sets of signals described before. It's important to note that this variance decomposition does not represent the relative importance of specific signals in forming expectations. Rather it should be interpreted as the relative importance of these signals to explain the time variation of marginal effects.

Table 4 shows the variance decomposition for time-varying marginal effects of two signals on expected economic conditions as presented in Section 4.2.2.[29] The top panel is for past/current signal on real GDP growth, denoted as $\beta^{Ey}_{y,t}$ and the bottom panel is for the future signal on real GDP growth (from SPF), denoted as $\beta^{Ey}_{Fy,t}$. In both marginal effects, signals on economic conditions contribute the most to the time-variation observed. They explain up to 57% of the variation for the marginal effect of the past signal and 52% for that of the future signal. News exposure to economic conditions also plays an important role, especially for the marginal effect of future signals. With signals and news exposure on economic conditions alone, I can explain as much as 72% and 80% of the total time-variation for the marginal effects of past and future

[29] For same decomposition exercise of unemployment expectations refer to Online Appendix D.4

signals.

Table 4: Variance Decomposition of Time-varying Marginal Effects: $E\Delta y$

Marginal Effect of Past Signal:		$\beta^{Ey}_{y,t}$				
Signal Type:		Economic Condition	Inflation	Interest rate	News	Total
	State $\theta_{i,t-1}$	17%	8%	3%	12%	40%
Channel:	Covariate $Z_{i,t}$	**40%**	12%	5%	3%	60%
	Total	57%	20%	8%	15%	
Marginal Effect of Future Signal:		$\beta^{Ey}_{Fy,t}$				
Signal Type:		Economic Condition	Inflation	Interest rate	News	Total
	State $\theta_{i,t-1}$	13%	2%	5%	9%	29%
Channel:	Covariate $Z_{i,t}$	**39%**	7%	6%	19%	71%
	Total	52%	9%	11%	28%	

On the other hand, inflation and interest rate signals account for only little of the time-variation, except for inflation signals in explaining marginal effects of past signal $\beta^{Ey}_{y,t}$. This is due to the signal on real oil price included as signals on inflation. Researchers document that oil price affects consumer expectations not only on inflation but also general economic conditions,[30] it is possible that oil prices either interact with or competing the attention put on signals about economic conditions and thus affecting the sensitivity of the household's expectation to these signals. Excluding oil price cuts down the marginal effect of Δy explained by inflation signals from 20% to 12%.

Another important question is for the same set of signals considered whether the time-variation of marginal effect is coming from contemporaneous signals $Z_{i,t}$ or through the accumulation of past signals which is represented by state $\theta_{i,t-1}$. I then separately evaluate the variation explained by these two channels. In Table 4 for each set of signals, I also document the variance explained by each channel separately. For economic condition signals, new information at each period plays the most important role, which is around 70% of the total variation explained by these signals. Meanwhile, the state also accounts for a significant share of the time-variation. It explains 17% and 13% respectively for the marginal effects of past and future signals. This means the weight households put on economic condition signals depends on not only their current level but also the state they accumulated from observing these signals in the past.

[30]See Edelstein and Kilian (2009), for example.

Variance decomposition shows that the most important signals for explaining time variation are those about economic conditions. But it does not offer information about how exactly these signals change marginal effects over time. It is possible that despite these signals being most important, they do not create the weight increase for future signals and decrease for past signals during bad times. To complete the picture, I present the time-varying marginal effects with only signals on economic conditions in Figure 4 and compare it with the actual marginal effects.

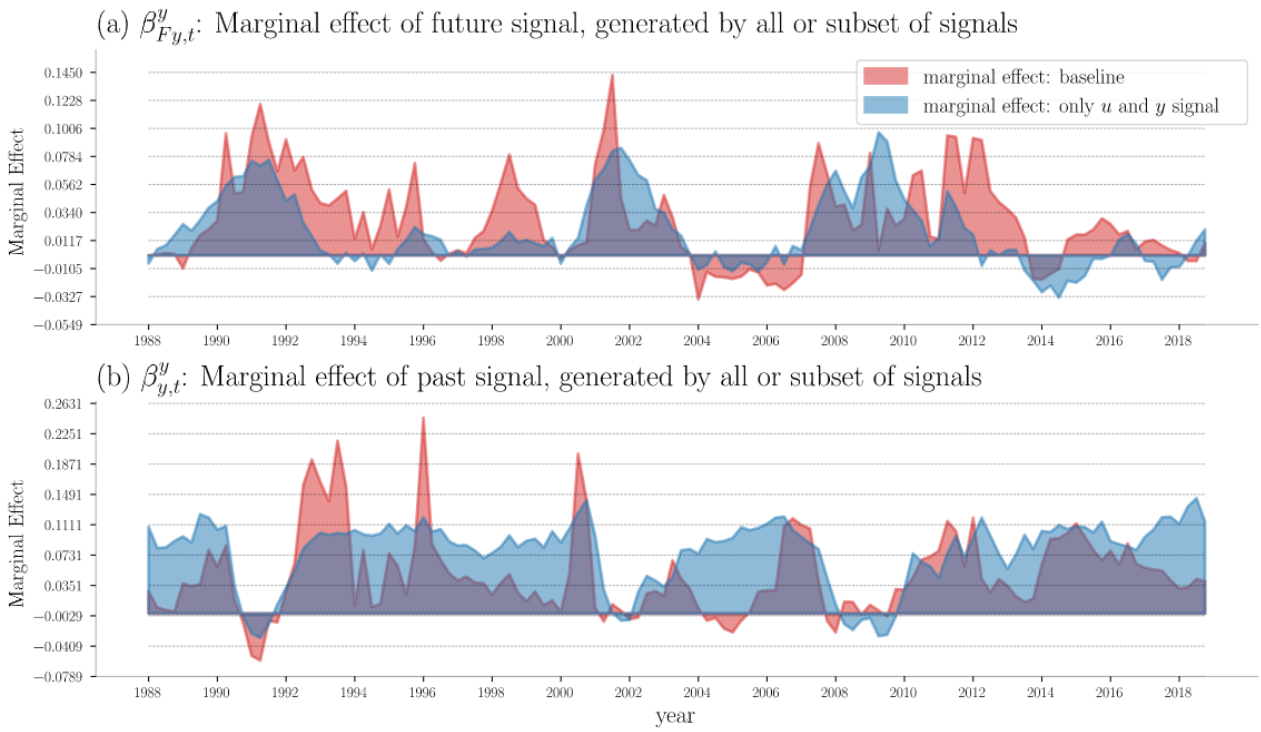


Figure 4: Time-varying marginal effect of past and future signal on real GDP growth. Top panel: marginal effect of future signal, $\beta^{Ey}_{Fy,t}$; bottom panel: marginal effect of future signal, $\beta^{Ey}_{y,t}$. The red curve: marginal effect created by estimated ASF with all signals. The blue curve: marginal effect created by ASF with only economic condition signals.

In Figure 4, the red curves are the baseline time-varying marginal effects from estimated ASF with all signals as input. The blue curves are marginal effects computed from ASF using only actual economic condition signals as input, which are the same series I use to perform variance decomposition in Table 4.[31] This figure shows strong evidence that economic condition signals generate the weight increase on future signals as well as a drop of weight on past signals during bad times. They are indeed key driving forces for the attention shift I documented before.

One other possible explanation for the time-variation of marginal effect was addressed by Carroll (2003), in which the author shows how information on inflation transmits from

[31] For signals other than economic conditions I use random draw from the empirical distribution of these signals.

professional forecasts to households through news media. Intuitively, when there are more news stories on economic conditions, it is easier for households to acquire information about the future, thus putting higher weights on those signals.[32] In the Online Appendix we include the same exercise as in Figure 4 but use either inflation and interest rate signals, or news exposures for robustness checks. Neither of these two types of information creates the attention shift pattern.

5 Model with Rational Inattention

In this section, I develop a simple two-period rational inattention model to illustrate how costly information acquisition alone can give rise to non-linear and state-dependent expectation formation as I documented in data. Comparing to a standard rational inattention model, as presented in Sims (2003) and Maćkowiak *et al.* (2018), several modifications are made to the model.

First, I allow agents to acquire information about both the current and future state of the economy, and there are two separate signals associated with this information. Such a modification is needed to address the attention-shift toward future signals. Secondly, rather than taking a linear-quadratic approximation of the agent's problem and looking for an analytical solution, I solve the problem numerically to keep the non-linear nature of the agent's optimal choices. This modification makes the value of information differs across states of the economy, which is the key mechanism to explaining the stylized facts documented in this paper.

5.1 Household's Problem

There is a representative household that faces an individual consumption-saving problem. The household lives for two periods and gets deterministic endowments $\{e_t, e_{t+1}\}$. The household can only save with a risky asset that pays a random return d_{t+1} at time $t+1$. The only uncertainty comes from d_{t+1}. I then interpret d_{t+1} as the fundamental economic condition in the future, as it accounts for all the uncertainty about the agent's future income.[33]

Before the agent chooses consumption and saving in the first period, he can obtain signals that help him to forecast d_{t+1}. After observing these signals, the agent forms a belief on the return of the risky asset and chooses consumption and saving according to this belief. In rational inattention models, the accuracy of signals is determined by the information structure. The agents can choose the information structure with a cost. Signals with high accuracy will have high costs. For now, I will denote the information structure chosen optimally by the agent as $\mathcal{I}_t$.

[32] See LAMLA and MAAG (2012) for example.

[33] If one considers saving as capital investment, with full depreciation d_{t+1} can be thought of as productivity shocks in the standard AK model.

The household's utility maximization problem then can be written as:

$$\max_{c_t, s_{t+1}} \quad \mathbb{E}[u(c_t) + \beta u(c_{t+1}) | \mathcal{I}_t] \tag{17}$$
$$s.t. \quad c_t + s_{t+1} = e_t$$
$$c_{t+1} = (1 + d_{t+1}) s_{t+1} + e_{t+1}$$

For ease of notation, define $r_{t+1} = 1 + d_{t+1}$ the above problem becomes:

$$\max_{s_{t+1}} \quad \mathbb{E}[u(e_t - s_{t+1}) + \beta u(r_{t+1} s_{t+1} + e_{t+1}) | \mathcal{I}_t] \tag{18}$$

5.2 Information Structure

For agents to make a forecast on d_{t+1}, I need to specify a law of motion for the stochastic return. Consider the return evolves according to an AR(1) process described in (19).

$$d_{t+1} = \rho d_t + \psi_{t+1} \tag{19}$$

To reflect the fact that there is information available to agents about the future of the fundamental, I assume the shock on return tomorrow has a predictable part η_t and an unpredictable part $\epsilon_{1,t+1}$. The predictable part itself follows a stationary AR(1) process.

$$\psi_{t+1} = \eta_t + \epsilon_{1,t+1} \tag{20}$$
$$\eta_t = \rho_\eta \eta_{t-1} + \epsilon_{2,t} \tag{21}$$

Both $\epsilon_{1,t+1}$ and $\epsilon_{2,t}$ are i.i.d and mean-zero shocks that follow normal distribution.[34] Denote $\boldsymbol{\epsilon}_t \equiv \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix} \sim N(\mathbf{0}, \boldsymbol{Q})$, $\boldsymbol{Q} \equiv \begin{bmatrix} \sigma^2_{1,\epsilon} & 0 \\ 0 & \sigma^2_{2,\epsilon} \end{bmatrix}$, $\boldsymbol{X}_t \equiv \begin{bmatrix} d_t \\ \eta_t \end{bmatrix}$, and $A \equiv \begin{bmatrix} \rho & 1 \\ 0 & \rho_\eta \end{bmatrix}$. I can write the state-space representation:

$$\boldsymbol{X}_{t+1} = A \boldsymbol{X}_t + \boldsymbol{\epsilon}_{t+1} \tag{22}$$

Signals: Because the model being analyzed here is not a Linear-Quadratic problem, the famous result that the optimal information set is Gaussian is not available. For simplicity, I restrict the signals considered here to be linear Gaussian. A convenient result of such restriction is that the choice of information set can be described by the precision (inverse of variances) of signals.

Before choosing the optimal information set with a cost, the household is also passively exposed to a signal on the current state d_t. This is summarized as a Gaussian noisy signal

[34] Such a formulation is similar to Barsky and Sims (2012), and the predictable part can be interpreted as "news shocks" described in Beaudry and Portier (2014). In general, this information may come from the stock market, news, or professionals. In this model, for simplicity I consider that this information is contained in the professional forecast. Throughout the model, I will assume the agent knows the correct law of motion of the stochastic return.

$z_0 = d_t + \xi_0$, where $\xi_0 \sim N(0, \sigma_z^2)$. Such a signal can be thought of as an information agent that picks up passively during daily life. The household's initial information set contains both her prior, $\boldsymbol{X}_0$, and the passive signal z_0. It can be fully summarized with updated prior: $\mathcal{I}_0 = \{\boldsymbol{X}_{t|0}\}$, with $\boldsymbol{X}_{t|0}$ stands for prior about $\boldsymbol{X}_t$ conditional on signal z_0.

Upon observing passive signals, agents also deliberately choose signals costly to be better informed. To be consistent with my empirical setup, I restrict the choices of signals to one about the current state d_t and one about the future that comes from SPF:

$$F_t d_{t+1} = \rho d_t + \eta_t \tag{23}$$

Agents observe unbiased signals on these two objects, with additive normal noise $\boldsymbol{\xi}_t$, where:

$$\boldsymbol{\xi}_t \equiv \begin{bmatrix} \xi_{1,t} \\ \xi_{2,t} \end{bmatrix}, \quad \boldsymbol{\xi}_t \sim N(\mathbf{0}, R), \quad R \equiv \begin{bmatrix} \sigma_{1,\xi}^2 & 0 \\ 0 & \sigma_{2,\xi}^2 \end{bmatrix}$$

Denote the vector of signals as $\boldsymbol{Z}_t$, the signal structure is given by:

$$\begin{bmatrix} z_t^{spf} \\ z_t \end{bmatrix} \equiv \boldsymbol{Z}_t = G\boldsymbol{X}_t + \boldsymbol{\xi}_t \tag{24}$$

Where G is given by $G = \begin{bmatrix} \rho & 1 \\ 0 & 1 \end{bmatrix}$. The information set after the agent chooses the precision of signals can be defined as $\mathcal{I}_t = \mathcal{I}_0 \cup \{\boldsymbol{Z}_t\}$.

Information Cost: Information comes with a cost. Following Sims (2003) I measure the cost of acquiring more information in set $\mathcal{I}_t$ with the difference of the Shannon entropy, denoted as $\mathcal{H}(.)$. As both random states and signals I considered are normally distributed, results from Maćkowiak *et al.* (2018) show that the entropy cost can be represented by posterior variance-covariance matrices. Denoted as κ, equation (25) formally defines the entropy cost.

$$\kappa = \mathcal{H}(\boldsymbol{X}_{t+1}|\mathcal{I}_0) - \mathcal{H}(\boldsymbol{X}_{t+1}|\mathcal{I}_t) = \frac{1}{2}log_2(\frac{det\Sigma_{t+1|0}}{det\Sigma_{t+1|t}}) \tag{25}$$

Where $\Sigma_{t+1|0}$ stands for posterior variance matrix for hidden states $\boldsymbol{X}_{t+1}$ conditional on information in $\mathcal{I}_0$ and $\Sigma_{t+1|t}$ stands for posterior variance matrix conditional on information in $\mathcal{I}_t$.[35]

5.3 Optimal Signals

Agent's problem comes in two steps. First, the agent chooses information set $\mathcal{I}_t$. He cannot control the realization of signal $\boldsymbol{Z}_t$ but he can choose the precision of noise $\boldsymbol{\xi}_t$ that is attached

[35] For derivations of entropy cost in (25) and the posterior variance-covariance matrices $\Sigma_{t+1|0}$ and $\Sigma_{t+1|t}$, please refer to the Online Appendix E.4.

to this signal. In this sense choosing information set $\mathcal{I}_t$ is equivalent to choosing variances of signal $\{\sigma^2_{1,\xi}, \sigma^2_{2,\xi}\}$. Then agent solves consumption-saving problem given the information set chosen and signals $\boldsymbol{Z}_t$ realized. This problem can be summarized as follows:

$$\max_{\sigma^2_{1,\xi},\sigma^2_{2,\xi}} \quad \mathbb{E}[u(e_t - s^*_{t+1}) + \beta u(r_{t+1}s^*_{t+1} + e_{t+1})|\mathcal{I}_0] - \lambda\kappa \tag{26}$$

$$s.t. \quad s^*_{t+1} = argmax_{s_{t+1}} \quad \mathbb{E}[u(e_t - s_{t+1}) + \beta u(r_{t+1}s_{t+1} + e_{t+1})|\mathcal{I}_t] \tag{27}$$

$$\kappa = \frac{1}{2}log_2(\frac{det\Sigma_{t+1|0}}{det\Sigma_{t+1|t}}) \tag{28}$$

The information cost in terms of utility loss is assumed to be a marginal cost parameter λ times the Shannon entropy cost κ. The parameter λ describes how costly it is for the agent to acquire information with some level of entropy reduction. When λ is bigger, it means the agent suffers higher utility loss from acquiring more information. In particular, when $\lambda = 0$, the information cost becomes irrelevant and the agent forms expectation according to FIRE.

For simplicity, assume quadratic utility function $u(c_t) = c_t - bc_t^2$.[36] The optimal saving conditional on information set from (27) is:

$$s^*_{t+1}(\mathcal{I}_t) = \frac{-1 + 2be_t + (\beta - 2b\beta e_{t+1})\mathbb{E}[r_{t+1}|\mathcal{I}_t]}{2b(1 + \beta\mathbb{E}[r^2_{t+1}|\mathcal{I}_t])} \tag{29}$$

Precisions of signals matter for the agent as they affect her optimal saving through $\mathbb{E}[r_{t+1}|\mathcal{I}_t]$ and $\mathbb{E}[r^2_{t+1}|\mathcal{I}_t]$.[37] Recall $r_{t+1} = 1 + d_{t+1}$, we have:

$$\begin{aligned}\begin{pmatrix}\mathbb{E}[d_{t+1}|\mathcal{I}_t]\\ \mathbb{E}[\eta_{t+1}|\mathcal{I}_t]\end{pmatrix} &= A\big((I - KG)\hat{\boldsymbol{X}}_{t|0} + K\boldsymbol{Z_t}\big)\\ &= A\Big((I - KG)\big((I - K_0G_0)\hat{\boldsymbol{X}}_0 + K_0z_0\big) + K\boldsymbol{Z_t}\Big)\end{aligned} \tag{30}$$

Where K is the Kalman Gain from signal $\boldsymbol{Z}_t$, $G_0 = \iota = [1 \quad 0]$, and K_0 the Kalman Gain from initial passive signal z_0. The expected second order term in the optimal saving function is then given by:

$$\mathbb{E}_t[r^2_{t+1}|\mathcal{I}_t] = \iota\Sigma_{t+1|t}\iota' + \big(1 + \mathbb{E}[d_{t+1}|\mathcal{I}_t]\big)^2 \tag{31}$$

From (30) and (31), the variances of signals, $\sigma^2_{1,\xi}$ and $\sigma^2_{2,\xi}$, affect the saving policy directly through the Kalman Filtering process and indirectly from random variable $\boldsymbol{Z}_t$. In general, a higher precision (or lower variance on the noise) leads to higher expected utility. More importantly, because the optimal saving choice is non-linear in the state, the ex ante expected

[36] Note that despite the utility function being quadratic, the problem doesn't boil down to an LQG as the policy function under full information is not linear in the state.

[37] For full derivation of $\mathbb{E}[r_{t+1}|\mathcal{I}_t]$ and $\mathbb{E}[r^2_{t+1}|\mathcal{I}_t]$, please refer to the Online Appendix E.5

utility in (26) depends not only on the variance of states conditional on $\mathcal{I}_0$ but also the mean of the states. This makes the expected benefit of information state-dependent.[38]

Finally, the information cost in (28) is also affected by signal precisions because the posterior variance-covariance matrix $\Sigma_{t+1|t}$ is given by:

$$\begin{aligned}\Sigma_{t+1|t} &= A\Sigma_{t|0}A' - AKG\Sigma_{t|0}A' + \boldsymbol{Q} \\ &= A\Sigma_{t|0}A' - A\Sigma_{t|0}G'(G\Sigma_{t|0}G' + R)^{-1}G\Sigma_{t|0}A' + \boldsymbol{Q}\end{aligned} \tag{32}$$

The trade-off agent faces in solving this problem are then between the benefit of more information and its cost. Lower $\sigma_{2,\xi}$ and $\sigma_{1,\xi}$ (thus higher precision on both signals of the current state and Professional Forecasts) will increase expected utility. Meanwhile, more accurate signals will also increase information cost κ, as accurate signals decrease the posterior variance of the agent's belief. Because the agent observes an initial signal z_0 which contains information about d_t, her optimal choice of signal precision will depend on d_t:[39] when d_t is negative, information becomes more valuable to the agent thus they are willing to choose higher precision for signals.

5.4 Results

Table 5: Model Parameters

Parameter	Value	Parameter	Value
e_t	10	e_{t+1}	5
b	1/40	β	0.95
ρ	0.2	ρ_η	0.9
$\sigma_{1,\epsilon}$	0.09	$\sigma_{2,\epsilon}$	0.09
σ_z	0.18	λ	0.042
$\hat{\boldsymbol{X}}_0$	$\mathbf{0}$		

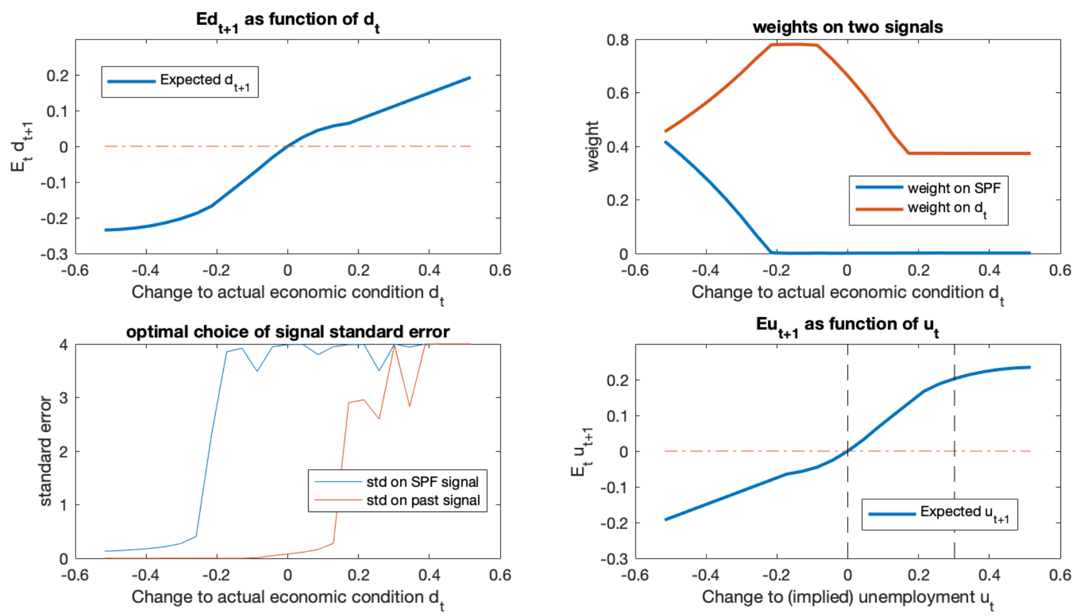
I solve the rational inattention problem (26)-(28) numerically using the parametrization included in Table 5. The main purpose of this section is to show that non-linear functional form and state-dependency weights can be generated with the proposed model with rational inattention.

[38]When the optimal saving is linear in states, the problem is a standard LQG problem where the expected benefit of information boils down to a form that only involves posterior variances and does not depend on the state. For a nice illustration please see the Online Appendix.

[39]If one assumes no passive signal is observed by agents, then the optimal choice of signal precision does not depend on d_t, but it will still depend on the prior belief about fundamentals. If this is the case, one should observe hidden states capturing most of the variation in time-varying marginal effects in Section 4.2.4. However, instead most variation is explained by the current signal, thus the empirical results are more consistent with the case when the agent observes a passive signal on current state d_t.

For direct comparison with my empirical finding, I first show counterfactual of expectation on d_{t+1} as a function of change to d_t, holding other signals at constant. I present it together with the agent's optimal choices of signal variances as well as the model implied weights on current (d_t) and future (SPF) signals. Recall the weights are computed directly from (30) using Kalman Filter. They are functions of model parameters as well as the endogenously chosen signals precisions. Specifically, the higher the precision on a signal, the higher the weight will be.[40] These results are included in Figure 5.

Figure 5: Results from Rational Inattention Model



Top left panel: expected state of economy Ed_{t+1} as function of current state d_t. Top right panel: red line is weight on past/current signal d_t, blue line is weight on future signal SPF. Bottom right panel: chosen standard deviation of noise attached to the corresponding signal. Red line is for signal on d_t, blue line is for signal on SPF. Bottom right panel: Implied expected unemployment as function of current unemployment. This is done by considering the unemployment state as the opposite of d_t. It is used to directly compare with Figure 2

Top left panel of Figure 5 can be seen as model implied Average Structural Function of agent's expectation formation process. It describes how expected future state $E_t d_{t+1}$ changes along the change of current state d_t. When realized d_t is high and positive, the slope of this function is quite flat. This is because agent believes it is more likely the state in future will be good, which indicates the return on risky asset is high in expectation. With this prior, more information is not valuable enough for agents thus they are not acquiring accurate signals on either current state d_t or SPF. This can be seen from bottom left panel: under this parametrization, any signal with noise variance higher than 1 implies almost 0 weight on this signal. When current state is good ($d_t > 0.2$) agent chooses variance on both signals to be

[40] I include the analytical derivation of the weights in the Online Appendix E.1.

higher than 10. The weight agent put on signal is depicted in top right panel. The reason why weight on d_t is not 0 is because of the initial signal on d_t that agent gets, before he chooses extra signals in the rational inattention model. This suggests when economic condition is good, agent will be happy to just form fuzzy expectation about future through the initial signal he gets, rather than actively searching for more information.

As the economic condition starts to get worse, in the area where $-0.2 < d_t < 0.2$, the slope of ASF gets most steep. This reflects the increasing weight agent puts on current signal about d_t. As agent realizes economic condition today is getting worse and worse (through observing the initial signal on d_t), information becomes more and more valuable and he is willing to pay higher cost to acquire more precise signals. This can be seen from bottom left graph that standard error on extra signals that agent chooses starts to fall sharply (which means precision of signal increases drastically) when current condition becomes worse. One interesting aspect is that they always get more accurate signal on d_t first before they go for SPF signal. This is because the information cost is increasing as agent's posterior getting more accurate. SPF signal contains more accurate information about future state thus is more costly for agents to get.

Finally when current economic condition is bad enough, when $d_t < -0.2$, agent gets more accurate signals on SPF. And because SPF has higher information content agent will start to put higher weights on signal about future (SPF) and lower weights on signal about current state d_t. Such a structure then created the non-linear ASF as I observed from survey data. Furthermore, it also generates the asymmetric response to good and bad states: as for positive realization of state d_t, agent has less incentive to acquire more information on it and end up attaching lower weights to the signal. This results in a lower mean expectation on d_{t+1}. On the other hand, when realization of d_t is bad, agent actively search for more information and put higher weights on these signals thus his expectation responds to bad states more than good ones.

The right bottom panel is then the ASF for implied unemployment expectation from the model. I consider $-d_t$ as a proxy for unemployment status because d_t can be interpreted as output growth and it is in general negatively correlated with unemployment. By doing this I can create the ASF for unemployment rate, which has the same dynamic as the one I found with RNN.

The time-variation of weights on signals is then reflected in top right panel of Figure 5: the weight on future signal (SPF) starts to increase when economic condition gets worse, meanwhile weight on past signals falls. To better illustrate this property of the model, I simulate the time series of d_t according to equations (19)-(21) for 200 periods.[41] Similar to the empirical part, I define episodes where d_t is 2 standard deviations lower than its mean as

[41] The bad periods account for 12 out of 200 periods of simulate d_t, which is similar to the recession periods as a fraction of post 1980 episode.

"bad periods". I then compute the average weight agent puts on past signal d_t and future signal $F_t d_{t+1}$, together with the optimal standard deviation of noise on each signal. Table 6 summarizes these statistics.

Table 6: Model Implied Weights and Precision during Bad and Ordinary Periods

	Bad Times		Ordinary Times	
Signal on:	Weights	Std. of noise	Weights	Std. of noise
Past/Current signal d_t	0.35	7e-4	0.57	0.94
Future signal $F_t d_{t+1}$	0.55	0.10	0.04	3.13

* Bad time is defined as periods in which d_t is 2 standard deviation lower from its long-run mean, 0. The rest episodes are considered as ordinary time. Weights are average model-implied weight on corresponding signal, during bad or ordinary time. Std. of noise is average model-implied standard deviation of noise on corresponding signal, during bad or ordinary time.

It is obvious in Table 6 that the model implies in the ordinary period, the agent will on average put higher weight on signals about past and current states when compared to bad times. The average weight on d_t is 0.57, almost twice as high as that when the economic condition is bad. Furthermore, the agent puts much higher weight on signals about the future during bad times, whereas almost no weight at all during ordinary times. The standard deviation of noise chosen by a rational inattentive agents then suggests such attention shift is induced by them optimally choosing much more accurate signals during bad times, whereas they choose to stay less informed during ordinary periods.

Finally, I want to point out that different values of information cost λ, prior mean and variances will also affect the state-dependency of information choices. I include these comparative statistic analyses in the Online Appendix.

6 Conclusion

How do households form expectations using a rich set of macroeconomic signals? This paper explores the answer to this question by proposing an innovative Generic Learning Framework that is flexible in functional forms and time-dependency that describe the relationship between signals and expectational variables. The unknown function form of the agents' expectation formation model is estimated with a Recurrent Neural Network. This method can recover any function forms considered by the Generic Learning Framework, including those most commonly used in the learning literature. After the functional estimation, I also obtain estimators on the average marginal effects of signals with valid inferences following the Double Machine Learning approach developed by Chernozhukov *et al.* (2018) .

Applying this method to survey data for US households, I document three stylized facts that are new to the literature: (1) agents' expectations about future economic conditions is a non-linear and asymmetric function of signals on real activities of the economy. (2) The attention to past and future signals in the Generic Learning Model is highly state-dependent. The agents behave like adaptive learners in ordinary periods and become forward-looking as the state of the economy gets worse. (3) Among all the signals considered in the empirical setup, signals on economic conditions play the most important role in creating the attention-shift. These findings are at odds with many models widely used in the literature, such as noisy information models and constant gain learning models.

Finally, a rational inattention model is developed to match these news stylized facts and help illustrate the impact of attention-shift on agents' expectation formation process. The model highlights that the agent's optimal choice of signal precision is a decreasing function of the current state of the economy due to non-linearity in their optimal saving choices. This information friction leads to the agent allocating more efforts to get information about the future when the economic condition deteriorates today. Such behavior makes them put higher weight on signals about the future and lower weight on information about current and past states. This information friction then is enough to generate both non-linear, asymmetric expectation and state-dependent weights on signals documented in the empirical findings.

References

Afrouzi, H. (2020). Strategic inattention, inflation dynamics, and the non-neutrality of money, cESifo Working Paper No. 8218, Available at SSRN: https://ssrn.com/abstract=3576296.

Andrade, P. and Le Bihan, H. (2013). Inattentive professional forecasters. *Journal of Monetary Economics*, **60** (8), 967–982.

Athey, S. (2018). The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An Agenda*, National Bureau of Economic Research, Inc, pp. 507–547.

— and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, **113** (27), 7353–7360.

Barsky, R. B. and Sims, E. R. (2012). Information, animal spirits, and the meaning of innovations in consumer confidence. *American Economic Review*, **102** (4), 1343–77.

Beaudry, P. and Portier, F. (2006). Stock prices, news, and economic fluctuations. *American Economic Review*, **96** (4), 1293–1307.

— and — (2014). News-driven business cycles: Insights and challenges. *Journal of Economic Literature*, **52** (4), 993–1074.

Bianchi, F., Ludvigson, S. C. and Ma, S. (2022). Belief distortions and macroeconomic fluctuations. *American Economic Review*, **112** (7), 2269–2315.

Blundell, R. and Powell, J. L. (2003). *Endogeneity in Nonparametric and Semiparametric Regression Models*, Cambridge University Press, *Econometric Society Monographs*, vol. 2, pp. 312–357.

Carroll, C. (2003). Macroeconomic expectations of households and professional forecasters. *The Quarterly Journal of Economics*, **118** (1), 269–298.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, **107** (5), 261–65.

—, —, —, —, —, — and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, **21** (1), C1–C68.

—, Newey, W. K. and Singh, R. (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica*, **90** (3), 967–1027.

Coibion, O. and Gorodnichenko, Y. (2012). What can survey forecasts tell us about information rigidities? *Journal of Political Economy*, **120** (1), 116 – 159.

— and — (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, **105** (8), 2644–78.

Cole, S. J. and Milani, F. (2020). *Heterogeneity in Individual Expectations, Sentiment, and Constant-Gain Learning*. Working Paper 8343, CESifo, Munich.

D'Acunto, F., Malmendier, U., Ospina, J. and Weber, M. (2020). Exposure to grocery prices and inflation expectations. *Journal of Political Economy*, **129** (5), 1615–1639.

Das, S., Kuhnen, C. M. and Nagel, S. (2019). Socioeconomic Status and Macroeconomic Expectations. *The Review of Financial Studies*, **33** (1), 395–432.

Edelstein, P. and Kilian, L. (2009). How sensitive are consumer expenditures to retail energy prices? *Journal of Monetary Economics*, **56** (6), 766–779.

Eusepi, S. and Preston, B. (2011). Expectations, learning, and business cycle fluctuations. *American Economic Review*, **101** (6), 2844–72.

Evans, G. W. and Honkapohja, S. (2001). *Learning and Expectations in Macroeconomics*. Princeton University Press.

Farmer, L. E. and Toda, A. A. (2017). Discretizing nonlinear, non-gaussian markov processes with exact conditional moments. *Quantitative Economics*, **8** (2), 651–683.

Farrell, M. H., Liang, T. and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, **89** (1), 181–213.

Flynn, J. P. and Sastry, K. (2022). Attention cycles. *Available at SSRN 3592107*.

Hamilton, J. (2016). Macroeconomic regimes and regime shifts. vol. 2, *Chapter 3*, Elsevier, pp. 163–201.

Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **2** (5), 359 – 366.

Hou, C. (2020). Uncovering subjective models from survey expectations. *Available at SSRN 3728884*.

Isakin, M. and Ngo, P. V. (2020). Variance decomposition analysis for nonlinear economic models. *The Oxford Bulletin of Economics and Statistics (forthcoming)*.

Kamdar, R. (2019). *The Inattentive Consumer: Sentiment and Expectations*. 2019 Meeting Papers 647, Society for Economic Dynamics.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, **105** (5), 491–95.

Lamla, M. J. and Lein, S. M. (2014). The role of media for consumers' inflation expectation formation. *Journal of Economic Behavior & Organization*, **106**, 62 – 77.

LAMLA, M. J. and MAAG, T. (2012). The role of media for inflation forecast disagreement of households and professional forecasters. *Journal of Money, Credit and Banking*, **44** (7), 1325–1350.

Maćkowiak, B., Matějka, F. and Wiederholt, M. (2018). Dynamic rational inattention: Analytical results. *Journal of Economic Theory*, **176**, 650 – 692.

Mackowiak, B. and Wiederholt, M. (2009). Optimal sticky prices under rational inattention. *American Economic Review*, **99** (3), 769–803.

Malmendier, U. and Nagel, S. (2015). Learning from Inflation Experiences *. *The Quarterly Journal of Economics*, **131** (1), 53–87.

Milani, F. (2007). Expectations, learning and macroeconomic persistence. *Journal of Monetary Economics*, **54** (7), 2065–2082.

PFAJFAR, D. and SANTORO, E. (2013). News on inflation and the epidemiology of inflation expectations. *Journal of Money, Credit and Banking*, **45** (6), 1045–1067.

Roth, C., Settele, S. and Wohlfart, J. (2020). Risk exposure and acquisition of macroeconomic information, available at SSRN: https://ssrn.com/abstract=3612751 or http://dx.doi.org/10.2139/ssrn.3612751.

Schäfer, A. M. and Zimmermann, H. G. (2006). Recurrent neural networks are universal approximators. In S. D. Kollias, A. Stafylopatis, W. Duch and E. Oja (eds.), *Artificial Neural Networks – ICANN 2006*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 632–640.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, **50** (3), 665 – 690, swiss National Bank/Study Center Gerzensee Conference on Monetary Policy under Incomplete Information.

— (2006). Rational inattention: Beyond the linear-quadratic case. *American Economic Review*, **96** (2), 158–163.

Sonoda, S. and Murata, N. (2015). Neural network with unbounded activations is universal approximator. *CoRR*, **abs/1505.03654**.

Woodford, M. (2001). Imperfect common knowledge and the effects of monetary policy. (8673).

Appendices

A Proof and Derivation

Proof of Theorem 1:

From (4), the average structural function is written as:

$$y_{i,t+1|t} \equiv \mathbb{E}_{\{\epsilon_{i,\tau}\}_{\tau=0}^{t}}[Y_{i,t+1|t}]$$

Under independence assumption 2, this is equivalent to counterfactual conditional expectation functions $\mathbb{E}[Y_{i,t+1|t}|\{Z_{i,\tau}\}_{\tau=0}^{t}]$:

$$\begin{aligned}
\mathbb{E}[Y_{i,t+1|t}|\{Z_{i,\tau}\}_{\tau=0}^{t}] &= \int F(\Theta_{i,t}) d\mathcal{F}_{\Theta_{i,t}}(\Theta_{i,t}|\{Z_{i,\tau}\}_{\tau=0}^{t}) \\
&= \int F(\Theta_{i,t}) \mathcal{P}_{\Theta_{i,t}}(\Theta_{i,t}|\{Z_{i,\tau}\}_{\tau=0}^{t}) d\Theta_{i,t} \\
&= \int \left(\int F(\Theta_{i,t}) \mathcal{P}_{\Theta_{i,t}}(\Theta_{i,t}|\{Z_{i,\tau}\}_{\tau=0}^{t}, \Theta_{i,t-1}) d\Theta_{i,t} \right) \mathcal{P}_{\Theta_{i,t-1}}(\Theta_{i,t-1}|\{Z_{i,\tau}\}_{\tau=0}^{t}) d\Theta_{i,t-1}
\end{aligned} \tag{33}$$

The first equality holds from Assumption 2. The conditional CDF of variable X is represented by $\mathcal{F}_X$ and conditional PDF is represented by $\mathcal{P}_X$. The third equality holds from Bayes Rule.

Now consider the conditional PDF $\mathcal{P}_{\Theta_{i,t}}(\Theta_{i,t}|\{Z_{i,\tau}\}_{\tau=0}^{t}, \Theta_{i,t-1})$, under assumption 2 it can be represented by PDF with respect to the i.i.d random variable $\epsilon_{i,t}$:

$$\begin{aligned}
\mathcal{P}_{\Theta_{i,t}}(\Theta_{i,t} = r'|\{Z_{i,\tau}\}_{\tau=0}^{t}, \Theta_{i,t-1}) &= \mathcal{P}_{\epsilon_{i,t}}(H(\Theta_{i,t-1}, Z_{i,t}, \epsilon_{i,t}) = r'|Z_{i,t}, \Theta_{i,t-1}) \\
&= \mathcal{P}_{\Theta_{i,t}}(\Theta_{i,t} = r'|Z_{i,t}, \Theta_{i,t-1})
\end{aligned} \tag{34}$$

Furthermore, as $\epsilon_{i,t}$ is i.i.d across time, this conditional probability is time-homogenous conditional on the same realization of $Z_{i,t}$:

$$\begin{aligned}
\mathcal{P}_{\Theta_{i,t}}(\Theta_{i,t} = r'|Z_{i,t} = z, \Theta_{i,t-1} = r) &= \mathcal{P}_{\epsilon_{i,t}}(H(r, z, \epsilon_{i,t}) = r') \\
&= \mathcal{P}_{\epsilon_{i,t+s}}(H(r, z, \epsilon_{i,t+s}) = r') \\
&= \mathcal{P}_{\Theta_{i,t+s}}(\Theta_{i,t+s} = r'|Z_{i,t+s} = z, \Theta_{i,t+s-1} = r) \quad \forall s > 0
\end{aligned} \tag{35}$$

Now one can discretize the continuous-state Markov Process.[42] Denote the grid points obtained for $\Theta_{i,t}$ as $D_r = \{x_r\}_{r=1}^{N_r}$ and corresponding transition probability from state r to r' as $\{p_{r,r'}(z)\}$. Now consider a finite dimensional variable:

$$\theta_{i,t}^{r} = \mathcal{P}_{\Theta_{i,t}}(\Theta_{i,t} = x_r|\{Z_{i,\tau}\}_{\tau=0}^{t}) \quad \forall r \in \{1, ..., N_r\}$$

[42] Following Farmer and Toda (2017), one can discretize non-linear non-Gaussian Markov Process and match exact conditional moments of the process, which is the same as my goal here. The details for the discretization procedure are included in Algorithm 2.2 from their paper.

Then it follows immediately from (33) that:

$$y_{i,t+1|t} = \mathbb{E}[Y_{i,t+1|t}|\{Z_{i,\tau}\}_{\tau=0}^{t}] = \sum_{r=1}^{N_r} F(x_r)\theta_{i,t}^{r} = f(\theta_{i,t})$$

Where the last equation is the definition of $f(.)$ function in theorem 1.

As $\theta_{i,t}$ is a function of history of signals $\{Z_{i,\tau}\}_{\tau}^{t}$, and it explicitly depends on $\theta_{i,t-1}$ as well as $Z_{i,t}$. This can be easily seen by induction, for $t = 0$:

$$\theta_{i,0}^{r} = \mathcal{P}_{\Theta_{i,0}}(\Theta_{i,0} = x_r|\Theta_{i,-1}, Z_{i,0})$$

For $t = 1$:

$$\begin{aligned}
\theta_{i,1}^{r'} &= \mathcal{P}_{\Theta_{i,1}}(\Theta_{i,1} = x_{r'}|\Theta_{i,-1}, Z_{i,0}, Z_{i,1}) \\
&= \sum_{r=1}^{N_r} \mathcal{P}_{\Theta_{i,1}}(\Theta_{i,1} = x_{r'}|\Theta_{i,0} = x_r, Z_{i,1} = z)\mathcal{P}_{\Theta_{i,0}}(\Theta_{i,0} = x_r|\Theta_{i,-1}, Z_{i,0}) \\
&= \sum_{r=1}^{N_r} p_{r,r'}(z)\theta_{i,0}^{r}
\end{aligned}$$

Where the second equality from above follows from Markov Property (34) then with time-homogeneity (35), one can get time t relation by induction:

$$\theta_{i,t}^{r'} = \sum_{r=1}^{N_r} p_{r,r'}(Z_{i,t})\theta_{i,t-1}^{r} \tag{36}$$

Equation (36) can be summarized as $\theta_{i,t} = h(\theta_{i,t-1}, Z_{i,t})$ from theorem 1. □

B Double De-biased Machine Learning Estimator

In this section I follow the semi-parametric moment condition model of Chernozhukov *et al.* (2018) and Chernozhukov *et al.* (2017). This is a general formulation that can be applied to estimation problems that involve:

- A finite dimensional parameter of interest – the average marginal effect defined in (7) β;
- Nuisance parameters that is usually infinite dimensional, denoted as η;
- Moment Condition that is (near) Neyman Orthogonal, denoted as $\mathbb{E}[\psi(W, \beta, \eta)]$, where $W = \{Y, X\}$ are the data observed;

I first focus to derive the Neyman Orthogonal Moment Condition for the estimation problem of average marginal effect. Throughout this appendix, denote $\ell(.)$ as objective function, g_t as average structural function that can be written as $g(\{X_{i,\tau}\}_{\tau=0}^{t}, \theta_{-1}) = f(h(X_{i,t}, \theta_{i,t-1}))$, $g_{t,x}^{j}$ as partial derivative of g_t with respect to j-th element of X, then $P(.)$ as the joint density function of input variables X. Suppose the true functional form of Average Structural Equation is $\mathbb{E}[Y_{i,t+1|t}|\{X_{i,\tau}\}_{\tau=0}^{t}] = g_{t,0}$ and the parameter of interest for each j-th element of the vector of average marginal effect $\mathbb{E}[\frac{\partial g_{t,0}}{\partial X^j}] = \beta_{t,0}^{j}$.

B.1 Neyman Orthogonal Moment Condition

1. Begin by declaring joint objective function, **at each time point** t, denote $X \equiv \{X_{i,\tau}\}_{\tau=0}^{t}$ for short-hand:

$$\min_{\beta_t^j, g_t} \quad \mathbb{E}[\ell(\{Y, X, \theta_{-1}\}; \beta_t, g_t)]$$

$$\ell(\{Y, X, \theta_{-1}\}; \beta_t, g_t) = 1/2(y - g_t(X, \theta_{-1}))^2 + \sum_j 1/2(\beta_t^j - g_{t,x}^j(X, \theta_{-1}))^2$$

Following Chernozhukov *et al.* (2018), the only requirement for objective function is the true value $g_{t,0}$ and $\beta_{t,0}^j$ for $\forall j$ minimize the objective function.

2. Concentrated-out non-parametric part:

$$g_{t,\beta_t} = argmin_g \mathbb{E}[\ell(\{Y, X, \theta_{-1}\}; \beta_t, g_t)]$$

Need to derive g_{t,β_t} using functional derivative. Notice:

$$\begin{aligned}\mathbb{E}[\ell(\{Y, X, \theta_{-1}\}; \beta_t, g_t)] &= \int \mathbb{E}[\ell()|X, \theta_{-1}]P(X, \theta_{-1})d(X, \theta_{-1}) \qquad (37)\\ &\equiv \int \mathcal{L}(\{X, \theta_{-1}\}; \beta_t, g_t, g_{t,x})d(X, \theta_{-1})\end{aligned}$$

Using Euler-Lagrangian Equation:

$$\begin{aligned}0 &= \frac{\partial \mathcal{L}}{\partial g_t} - \sum_j^J \frac{\partial}{\partial x_t^j}\Big(\frac{\partial \mathcal{L}}{\partial g_{t,x}^j}\Big) \qquad (38)\\ &= \underbrace{-(\mathbb{E}[Y|X, \theta_{-1}] - g_t(X, \theta_{-1}))P(X, \theta_{-1})}_{\equiv \frac{\partial \mathcal{L}}{\partial g_t}} - \sum_j^J \frac{\partial}{\partial x_t^j} \underbrace{\big(-(\beta_t^j - g_{t,x}^j(X, \theta_{-1}))P(X, \theta_{-1})\big)}_{\equiv \frac{\partial \mathcal{L}}{\partial g_{t,x}^j}}\\ &= -(\mathbb{E}[Y|X, \theta_{-1}] - g_t(X, \theta_{-1}))P(X, \theta_{-1}) +\\ &\sum_j^J \big(-g_{t,xx}^j(X, \theta_{-1})P(X, \theta_{-1}) + \frac{\partial P(X, \theta_{-1})}{\partial x_t^j}(\beta_t^j - g_{t,x}^j(X, \theta_{-1}))\big)\end{aligned}$$

The concentrated-out non-parametric part at time t then is given by:

$$g_{t,\beta_t}(X, \theta_{-1}) = \mathbb{E}[Y|X, \theta_{-1}] + \sum_j^J \big(g_{t,xx}^j(X, \theta_{-1}) - \frac{\partial ln[P(X, \theta_{-1})]}{\partial x_t^j}(\beta_t^j - g_{t,x}^j(X, \theta_{-1}))\big)$$

3. Concentrated Objective at each time t:

$$\min_{\beta_t} \mathbb{E}[1/2(Y - g_{t,\beta_t}(X, \theta_{-1}))^2 + \sum_j 1/2(\beta_t^j - g_{t,\beta_t,x}^j(X, \theta_{-1}))^2]$$

Take F.O.C with respect to β_t^j and evaluate at $g_{t,\beta_t} = g_{t,0}$:

$$\mathbb{E}[\beta_t^j - g_{t,0,x}^j(X, \theta_{-1}) + \frac{\partial ln(P(X, \theta_{-1}))}{\partial x^j}(Y - g_{t,0}(X, \theta_{-1}))] = 0$$

Now notice two things here:

- In this set-up basically at each time t the $g_{t,0}()$ function is different, so that β_t is different as well. Without proper regularity the g function could be non-stationary. This is when the markov assumptions come to play. The assumptions with $f()$ and $h()$ functions basically interpret the time-varying β_t is because of different states θ_{t-1}.
- With the previous approach, we get moment condition of β_t^j instead of β^j,they are different because (1) $g_{t,0}(.)$ function is different at each t; (2) $P(X, \theta_0)$ is changing at each t.

  The first problem is solved by the Markov property and hidden variable:

$$g_{t,0}(\{X_{i,\tau}\}_{\tau=0}^t, \theta_{i,-1}) \equiv \mathbb{E}[Y_{i,t+1|t}|\{X_{i,\tau}\}_{\tau=0}^t, \theta_{i,-1}] = f(\theta_{i,t}) = f(h(\theta_{i,t-1}, X_{i,t}))$$

  Plug this into the moment condition:

$$\mathbb{E}[\beta_t^j - f \circ h_{x^j}(\theta_{i,t-1}, X_{i,t}) + \frac{\partial ln(P(X, \theta_{-1}))}{\partial x^j}(Y - f \circ h(\theta_{i,t-1}, X_{i,t})] = 0 \tag{39}$$

  The second problem can be solved by assuming dependency of $X_{i,t}$ and $X_{i,t-s}$. As θ_{-1} are assumed to be zeros in practice, which is deterministic. Here I assume variables $X_{i,t}$ follow a VAR(1) process so that:

$$\begin{aligned} P(X_{i,t}, X_{i,t-1}, ..., X_{i,0}) &= P(X_{i,t}|X_{i,t-1}, ..., X_{i,0})P(X_{i,t-1}|X_{i,t-2}, ..., X_{i,0})...P(X_{i,0}) \\ &= P(X_{i,t}|X_{i,t-1})P(X_{i,t-1}|X_{i,t-2})...P(X_{i,0}) \end{aligned}$$

  This leads to the fact that $\frac{\partial ln(P(X_{i,t}, X_{i,t-1}))}{\partial X_{i,t}^j} = \frac{\partial ln(P(X_{i,t}, X_{i,t-1}, ... X_{i,0}))}{\partial X_{i,t}^j}$. For this reason, in practice, I just need to estimate the joint density function $P(X_{i,t}, X_{i,t-1})$. Then equation (39) leads to moment condition (9) given the fact that $g(\{X_\tau\}_{\tau=0}^t, \theta_{-1}) \equiv f \circ h(\theta_{i,t-1}, X_{i,t})$.

B.2 Verifying Moment Condition is Orthogonal

This can be done by computing the Frechet Derivative with respect to nuisance parameter g of the moment condition $\mathbb{E}[\psi(W, \beta, \eta)]$, notice that $\eta = \{g, P\}$. The estimate of g will later be obtained from RNN. For sake of simplified notation, I drop the t and consider 1 dimensional case, but the appliction can be easily extended to multidimensional case.

$$\psi(W, \beta, \eta) \equiv \beta - g'(X) + \frac{P'(X)}{P(X)}(\mathbb{E}[Y|X] - g(X)) \tag{40}$$

Define functional $F : C(\mathbb{R}) \rightarrow C(\mathbb{R})$:

$$F(g)(\beta, X) = \mathbb{E}[\psi(W, \beta, \eta)]$$

The Frechet Derivative along direction v is given by:

$$\begin{aligned}
F(g+v) - F(g) &= \mathbb{E}[-v'(X) - \frac{P'(X)}{P(X)}v(X)] \qquad (41)\\
&= \lim_{\delta \rightarrow 0} \mathbb{E}[-\frac{v(X+\delta) - v(X)}{\delta} - \frac{P(X) - P(X-\delta)}{P(X)\delta}v(X)]\\
&= \lim_{\delta \rightarrow 0} 1/\delta[-\int_X v(X+\delta)P(X)dX + \int v(X)P(X)dX\\
&- \int v(X)P(X)dX + \int v(X)P(X-\delta)dX]\\
&= \lim_{\delta \rightarrow 0} 1/\delta[\int_y v(y+\delta)P(y)dy - \int_x v(x+\delta)P(x)dx]\\
&= 0
\end{aligned}$$

B.3 High Level Assumptions on Nuisance Parameters

To ensure the asymptotic property of estimate $\hat{\beta}$ obtained from DML approach to hold, I refer to Theorem 3.1 from Chernozhukov *et al.* (2018). First denote the moment condition derived in **Appendix B.1** as $\psi(W, \beta, \eta)$, where β is the parameter of interest, X is data in use and $\eta = \{g, P\}$ are nuisance parameters estimated from functional estimation, where $g(.)$ is ASF and $P(.)$ is joint density function of X. To apply this theorem one needs to verify three condition[43]:

1. Moment condition(scores) is linear in parameter of interest, β:

$$\psi(W, \beta, \eta) = \psi^a(W, \eta)\beta + \psi^b(W, \eta)$$

2. (Near) Neyman Orthogonality of score $\psi(W, \beta, \eta)$;

3. Fast enough convergence of nuisance parameters $\eta = \{g, P\}$. Notice such condition is formally described by Assumption 3.2 in Chernozhukov *et al.* (2018). And the authors discussed the sufficient conditions for this assumption to hold: ψ is twice differentiable and $\mathbb{E}[(\hat{\eta}(X) - \eta_0(X))^2]^{1/2} = o(n^{-1/4})$. And the variance of score ψ, $\mathbb{E}[\psi(W, \beta, \eta)\psi(W, \beta, \eta)']$ is non-degenerate.

Condition 1 is obvious given the Neyman Orthogonal score derived in Appendix B.1: equation (40) is linear in β. Condition 2 is verified in Appendix B.2.

[43] In Chernozhukov *et al.* (2018) these conditions are defined formally by their Assumption 3.1 and 3.2.

The convergence speed requirement in condition 3 needs a bit of work. In practice $g(.)$ function will be estimated by RNN and $P(.)$ function is estimated with gaussian kernel density estimation. For RNN the convergence speed of estimate $\hat{g}$ is offered by Theorem 1 of Farrell *et al.* (2021). To achieve the convergence speed described there, one needs to put restrictions on width and depth of neural network used to approximate $g(.)$. Specifically, for input dimension d, sample size n and smoothness of function $g(.)$, θ, one needs width $H \asymp n^{\frac{d}{2(\theta+d)}}$ and depth $L \asymp \log n$. These conditions will guarantee a convergence speed on a level of $\{n^{-\theta/(\theta+d)\log^8 n + \frac{\log\log n}{n}}\}$ which is faster than $n^{-1/2}$.[44] My baseline architecture satisfies these restrictions.

For convergence speed of joint density $P(.)$, it is estimated by gaussian kernel density estimation with Silverman Rule of Thumb for bandwidth selection. Denote the order of gaussian kernel as ν, and the input dimension of density function $P(.)$ as d' the asymptotic mean integrated squared error (AMISE) is known to be $O(n^{-2\nu/(2\nu+d')})$. The convergence speed requirement in condition 3 needs $2\nu/(2\nu+d') > 1/2$, or $\nu > d'/2$.[45] Notice the density function here is a joint density for $X_{i,t}$ and $X_{i,t-1}$ so its dimensionality is typically twice of the input for RNN. I then need to use a higher order gaussian kernel with at least $\nu = 28$ to ensure the convergence speed requirement for the density estimator.

Finally, after verifying all three pre-conditions, according to Theorem 3.1 from Chernozhukov *et al.* (2018), denoting the Jacobian matrix from the Neyman Orthogonal score as J_0 and the true value of nuisance parameter as η_0:

$$J_0 = \mathbb{E}[\psi^a(W, \eta_0)]$$

The DML estimator $\hat{\beta}$ is then centered at true values β_0 and are approximately linear and Gaussian:

$$\sqrt{n}\sigma^{-1}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i}^{n} \bar{\psi}(W_i) \to N(0, I_d)$$

Where $\bar{\psi}(.)$ is the influence function of the form:

$$\bar{\psi}(W_i) = -\sigma^{-1} J_0^{-1} \psi(W_i, \beta_0, \eta_0)$$

The σ^2 is the variance that is given by:

$$\sigma^2 = J_0^{-1}\mathbb{E}[\psi(W_i, \beta_0, \eta_0)\psi(W_i, \beta_0, \eta_0)'](J_0^{-1})'$$

Notice in my case $\psi^a(W, \eta_0) = 1$ so that $J_0 = 1$. This is the same distribution as if we plug in the true nuisance parameters η_0 and is enough for asymptotic inferences.

[44] See Theorem 1 in Farrell *et al.* (2021) for details.

[45] See *Hansen (2009)* for details

Online Appendix for "Learning and Subjective Expectation Formation: A Recurrent Neural Network Approach"

Chenyu(Sev) Hou

May 16, 2023

Appendices

C Examples

C.1 Example: Standard Noisy Information Model in Generic Learning Framework

In this subsection I take the standard noisy information model as an example and show how it can be represented by the Generic Learning Framework. The purpose of this example are three folds. First it gives an example of essential elements in the Generic Learning Model including hidden states $\Theta_{i,t}$, Average Structural Function and the transformed dynamic system (6) in the context of a familiar learning model. Secondly it illustrates how RNN performs in approximating the ASF (in this case linear) and estimating marginal effect without knowledge of the exact functional form of learning model. Lastly as I consider a special case when the expectation formation structure is still linear but OLS is mis-specified and show the performance of RNN in estimating the average marginal effect. This exercise illustrates the possible improvement in using RNN even in a linear case.

*I'm grateful to Jesse Perla, Paul Schrimpf, Paul Beaudry, Michael Devereux, and Amartya Lahiri for their invaluable guidance and support on this project. I thank Vadim Marmer, Fabio Milani, Monika Piazzesi, Henry Siu, Hassan Afrouzi, and many others for their insightful comments. I'm also thankful to the computational support of Compute Canada (www.computecanada.ca). All the remaining errors are mine.

†Chenyu Hou: Vancouver School of Economics, University of British Columbia. Email: sevhou1989@gmail.com

Data Generating Process: Consider agents want to predict inflation one period from now denoted as $\pi_{i,t+1|t}$. At time t, they can observe two signals $\{\pi_{i,t}, s_{i,t}\}$. There are two latent variables $\{\pi_t, L_t\}$ that they need to make inference of to form expectation of inflation. Represent the Actual Law of Motion as a Gaussian Linear State Space Model:

$$\begin{bmatrix} \pi_t \\ L_t \end{bmatrix} \equiv X_t = \boldsymbol{A} X_{t-1} + \epsilon_t \tag{42}$$

Where $\boldsymbol{A}$ describes how latent states X_t evolves along time, ϵ_t is i.i.d shock each period. Assume for simplicity the agent's Perceived Law of Motion is the same as (42). Agents do not observe X_t directly, instead they observe a noisy signals about it. Their observational equation is:

$$\begin{bmatrix} \pi_{i,t} \\ s_{i,t} \end{bmatrix} \equiv O_{i,t} = \boldsymbol{G} X_t + \nu_{i,t} \tag{43}$$

Both shock ϵ_t and $\nu_{i,t}$ are i.i.d and follow normal distribution with covariance matrix R and Q:

$$\epsilon_t \sim N(0, R) \quad \nu_{i,t} \sim N(0, Q)$$

This describes the standard noisy information model with two latent states. They use a stationary Kalman Filter to form prediction of the latent variable $X_{i,t+1|t}$, where $\boldsymbol{K}$ is the Kalman Gain.

$$\begin{bmatrix} \pi_{i,t+1|t} \\ L_{i,t+1|t} \end{bmatrix} \equiv X_{i,t+1|t} = \boldsymbol{A}(X_{i,t|t-1} + \boldsymbol{K}(O_{i,t} - \boldsymbol{G} X_{i,t|t-1})) \tag{44}$$

The Generic Learning Formulation The stationary Kalman Filter is a special case of Generic Learning Model. First notice the i.i.d error $\nu_{i,t}$ satisfies assumption 2. The expectation is also formed by filtering step and updating step:

$$X_{i,t|t} = X_{i,t|t-1} + \boldsymbol{K}(O_{i,t} - \boldsymbol{G} X_{i,t|t-1}) \quad \text{(Filtering Step)}$$

$$X_{i,t+1|t} = \boldsymbol{A} X_{i,t|t} \quad \text{(Forecasting Step)}$$

Replace $X_{i,t+1|t}$ with $\hat{Y}_{i,t+1|t}$ and define the "now-cast" variable $X_{i,t|t}$ as latent state variable $\Theta_{i,t}$ in Generic Learning Model, we can re-write Kalman Filter (44) as equation (45) and (46), which reflect the generic formulation of updating step (2) and forecasting step (3). It is obvious that in the stationary Kalman Filter case, both $F(.)$ and $H(.)$ are linear.

$$\hat{Y}_{i,t+1|t} = \boldsymbol{A}\Theta_{i,t} \tag{45}$$

$$\Theta_{i,t} = (\boldsymbol{A} - \boldsymbol{KGA})\Theta_{i,t-1} + \boldsymbol{KG} X_t + \boldsymbol{K}\nu_{i,t} \tag{46}$$

Average Structural Function I then turn to the ASF implied by Kalman Filter (45) and (46). This is simply done by taking expectation of $\hat{Y}_{i,t+1|t}$ conditional on observables X_t. The goal is to integrating out the i.i.d noise term $\nu_{i,t}$ which is not observable by econometrician. Now we can define the sufficient statistics for $\Theta_{i,t}$ as:

$$\theta_{i,t} = \mathbb{E}[\Theta_{i,t}|\{X_\tau\}_{\tau=0}^{t}] \tag{47}$$

Taking the expectation of (45) and (46) conditional on history of the observable $\{X_\tau\}_{\tau=0}^{t}$ it immediately follows:

$$\begin{aligned} y_{i,t+1|t} \equiv \mathbb{E}[\hat{Y}_{i,t+1|t}|\{X_\tau\}_{\tau=0}^{t}] = \boldsymbol{A}\theta_{i,t} \\ \theta_{i,t} = (\boldsymbol{A} - \boldsymbol{KGA})\theta_{i,t-1} + \boldsymbol{KG}X_t \end{aligned}$$

This illustrates the link between ASF with the underlying expectation formation model: in the linear case with mean zero error $\nu_{i,t}$, the function form from ASF, $f(.)$ and $h(.)$ are linear and are identical to those from the underlying expectation formation model.

Estimation with Simulated Sample Now suppose as econometricians we want to estimate marginal effect of two signals $\{\pi_t, s_{i,t}\}$ on $\pi_{i,t+1|t}$. The standard approach is to directly estimate the reduced-form equation derived from (44) with OLS. This requires $X_{i,t+1|t}$ observed for each t and the learning model is correctly specified. However in reality it is possible that expectation on latent state $L_{i,t+1|t}$ is not observable or not considered in the model[1]. If this is the case OLS with only lag term $\pi_{i,t|t-1}$ is included in the regression suffers from omitted variable problem.

On contrary, estimation with RNN does not require a correct specification on latent variable $\Theta_{i,t}$, and it doesn't need $L_{i,t|t-1}$ to be observable at all. To show this I simulated 100 random samples according to the Kalman Filter as in (44). In this experiment I consider three different models to estimate marginal effect of the two signals $\{\pi_t, s_{i,t}\}$: (1) the RNN with sequence of $\{\pi_\tau, s_{i,\tau}\}_{\tau=0}^{t}$ and lag expected inflation $\pi_{i,t|t-1}$ as input[2]; (2) mis-specified OLS that uses the same set of variables as dependent variable, the OLS is mis-specified because $L_{i,t+1|t}$ is not available to econometricians; (3) correctly specified OLS with $L_{i,t+1|t}$ observable, which is typically not available. I'll show RNN can still recover the linear relationship between signal

[1]For example, when agent form expectation on inflation, if they believe in a three equation New Keynesian Model, they may also want to infer demand and supply shocks as unobserved states. In a Kalman Filter that takes only inflation as unobserved state, OLS will suffer from omitted variable problem.

[2]Interestingly, for estimating ASF and marginal effect, one do not need to include the lag expectation $\pi_{i,t|t-1}$ in RNN, only history of signals are sufficient. The results without lag expectation are similar to these results I include here.

and expectational variable as well as obtain comparable estimate on signals as the correctly specified OLS estimator (BLUE in this case), whereas mis-specified OLS is heavily biased.

I first depict the recovered average structural function between inflation expectation $\pi_{i,t+1|t}$ and signals π_t, $s_{i,t}$ in Figure 6. The red solid line is the true Average Structural Function implied by the Kalman Filter (44) and the black solid line is the mean of estimated ASF from 100 random samples using RNN. I also plot estimated ASF for each sample in grey color. The top panel in Figure 6 is the ASF along dimension of realized inflation π_t and the bottom panel is along signal $s_{i,t}$. It is obvious that the estimated ASF all indicate linear relationship between signals and expected inflation. This means RNN will recover a linear function if the underlying expectation formation model is indeed linear. It also shows the stability of the performance of RNN: with 100 random samples it recovers the ASF relative close to the truth.

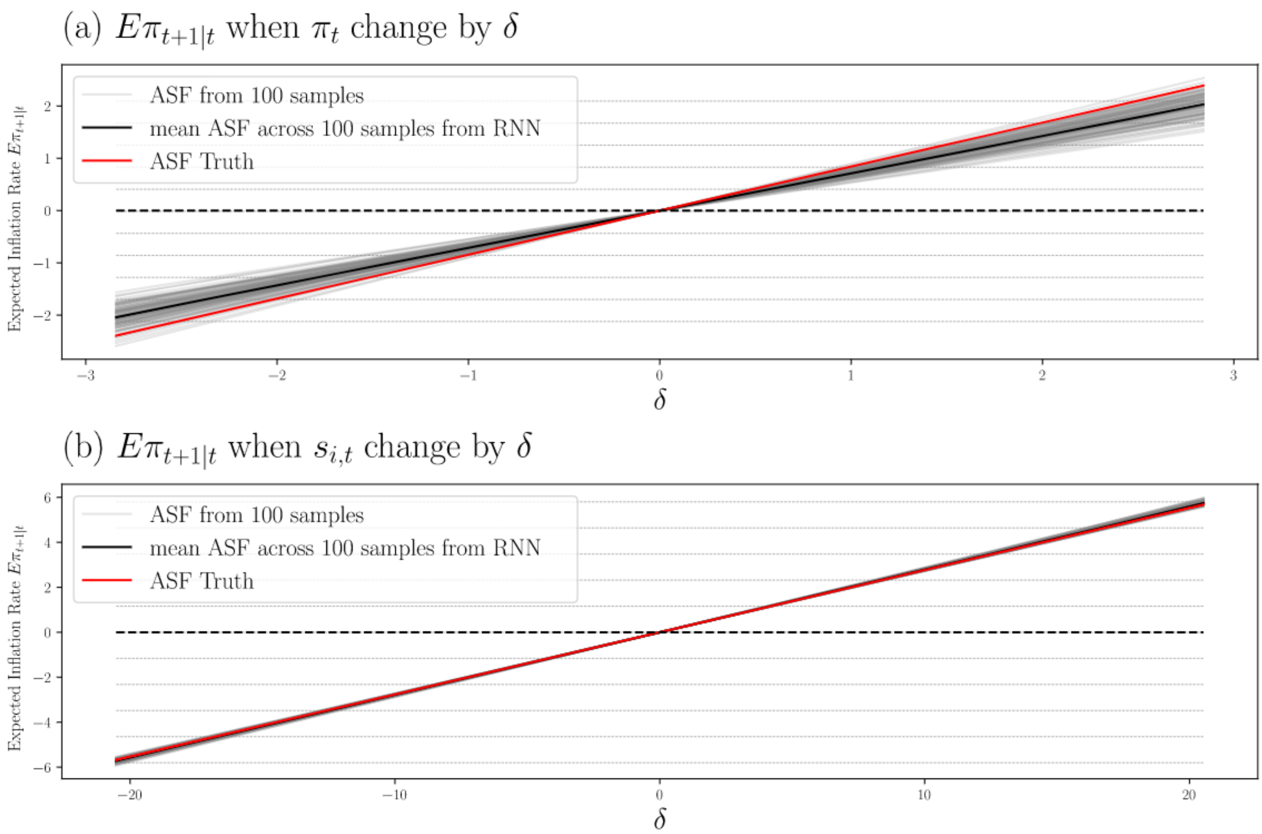


Figure 6: Estimated Average Structural Function from random samples using RNN. Function depicts change of expected variable in response to corresponding signal change by δ. Panel (a): expected inflation as function of inflation signal π_t. Panel (b): expected inflation as function of private signal $s_{i,t}$. Red solid line is the actual ASF implied by linear Kalman Filter. Solid black line is the mean of estimated ASF from 100 random samples. Grey lines are estimated ASFs from each random sample.

I then report the (naive) estimates of marginal effects from RNN and compare them to those from the other two models considered. The following table shows the estimation result from RNN, mis-specified OLS and correctly specified OLS. In this table, the first column is mean squared error on the whole sample, the second column is estimated marginal effect on signal π_t and third column is estimated marginal effect on signal $s_{i,t}$. In brackets I report the standard deviation of the estimate using 100 simulated random samples. Not surprisingly, correctly specified OLS is BLUE in this case with unbiased estimates and small standard deviations. However the key thing to notice here is that mis-specified OLS is biased due to the omitted latent state, whereas RNN has result that is consistent with the true marginal effect, with acceptable standard deviations across 100 samples.

Table 7: Performance of RNN v.s. OLS

	MSE	π_t	$s_{i,t}$
(1) RNN	2.91	0.82	0.276
	(0.054)	(0.037)	(0.003)
(2) OLS mis-specified	3.296	0.720	0.279
	(0.023)	(0.033)	(0.001)
(3) OLS correct	2.835	0.841	0.277
	(0.014)	(0.005)	(0.001)
Truth		0.842	0.277

* The first column is mean squared error on the whole sample, the second column is estimated marginal effect on signal π_t and third column is estimated marginal effect on signal $s_{i,t}$. In brackets I report the standard deviation of the statistics using 100 simulated random samples.

C.2 Example: Constant Gain Learning in Generic Learning Framework

In this subsection I will illustrate how a standard Constant Gain Learning model can be analytically expressed in the form of the Generic Learning Framework. An example of such model is from Evans and Honkapohja (2001). For simplicity I consider the one dimensional case, where an agent observes realized inflation π_t at each time t and try to form forecast about π_{t+1}. I also drop the individual indicator i to same some notations, but the framework can be easily generalized to multi-dimensional multi-agent case. The agent believes in a "Perceived Law of Motion" (PLM) about how inflation is evolving in time and try to estimate the relevant parameters in the PLM using observed data. To do this, she will run OLS at every period and apply a constant weight to the newly available data. With this learning scheme agent perceives different values for parameters in their PLM and form expectation accordingly. The model features a constant gain γ, which represents the weight the agent put on newly observed data. Let's assume the PLM the agent believes in is an AR(1) process:

$$\pi_{t+1} = b_0 + b_1\pi_t + \eta_{t+1} \quad \text{(PLM)}$$

In this setup, the parameters agent try to learn from realized data are b_0 and b_1. η_{t+1} stands for the mean zero i.i.d random shock realized in each period. The agent uses and OLS method to estimate b_0 and b_1 every period, and this process can be formulated recursively such that in each period the agent forms a different estimate b_t:

$$b_t = b_{t-1} + \gamma R_t^{-1}\boldsymbol{X}_{t-1}(\pi_t - b'_{t-1}\boldsymbol{X}_{t-1})$$
$$R_t = R_{t-1} + \gamma(\boldsymbol{X}_{t-1}\boldsymbol{X}'_{t-1} - R_{t-1})$$

$$\boldsymbol{X}_t = [1 \quad \pi_t]' \quad b = [b_0 \quad b_1]'$$

At time t, the agent then forms expectation about future inflation using the PLM, with some i.i.d noise attached on top of the endogenous component that comes from constant gain learning process, ϵ_t. This exogenous component is sometimes interpreted as "sentiment", for example in Cole and Milani (2020).

$$E_t\pi_{t+1} = b'_t\boldsymbol{X}_t + \epsilon_t \tag{48}$$

Now suppose the agent is learning with the above set-up. As observers we see: $\boldsymbol{X}_t, E_t\pi_{t+1}$ up to each time t. We do not see the hidden variables such as b_t and R_t. We also don't know the function form that connects the hidden variables, observables and expectational variables. The goal now is to represent the system described by this constant gain learning model in terms of the Generic Learning Framework. Define the hidden states $\Theta_t = [\boldsymbol{X}_t, b_t, R_t, \epsilon_t]'$. The recursive mapping from observables (and previous hidden states) to hidden states $H(.)$ then can be given by:

$$\Theta_t = H(\boldsymbol{X}_t, \Theta_{t-1}, \epsilon_t)$$

Where

$$\boldsymbol{X}_t \equiv H_1(\boldsymbol{X}_t, \Theta_{t-1}, \epsilon_t) = \boldsymbol{X}_t$$
$$R_t \equiv H_2(\boldsymbol{X}_t, \Theta_{t-1}, \epsilon_t) = R_{t-1} + \gamma(\boldsymbol{X}_{t-1}\boldsymbol{X}'_{t-1} - R_{t-1})$$
$$b_t \equiv H_3(\boldsymbol{X}_t, \Theta_{t-1}, \epsilon_t) = b_{t-1} + \gamma R_t^{-1}\boldsymbol{X}_{t-1}(\pi_t - b'_{t-1}\boldsymbol{X}_{t-1})$$
$$\epsilon_t \equiv H_4(\boldsymbol{X}_t, \Theta_{t-1}, \epsilon_t) = \epsilon_t$$

Notice here, as Θ_t can be any measurable function of $\boldsymbol{X}_t$, Θ_{t-1} and ϵ_t, it can certainly contain elements such as the input $\boldsymbol{X}_t$. Although $\boldsymbol{X}_t$ is actually observable, it remains "hidden" to econometrician as without further knowledge on expectation formation process, one will not know what the exact mapping from observables to elements of Θ_t is. Then the expectation formation model $F(.)$ is given by:

$$E_t\pi_{t+1} \equiv F(\Theta_t) = b_t'\boldsymbol{X}_t + \epsilon_t$$

Now I show that the expectation formed by constant gain learning can be analytically represented by the Generic Learning Framework described by updating step (2) and forecasting step (3). The Average Structural Function implied by this setup is straight forward: one can define $\theta_t = [\boldsymbol{X}_t, b_t, R_t]'$ and obtain $f(.)$ and $h(.)$ by integrating out the i.i.d random variable ϵ_t.

D Appendix on Empirical Findings

D.1 More on Time-varying Marginal Effect

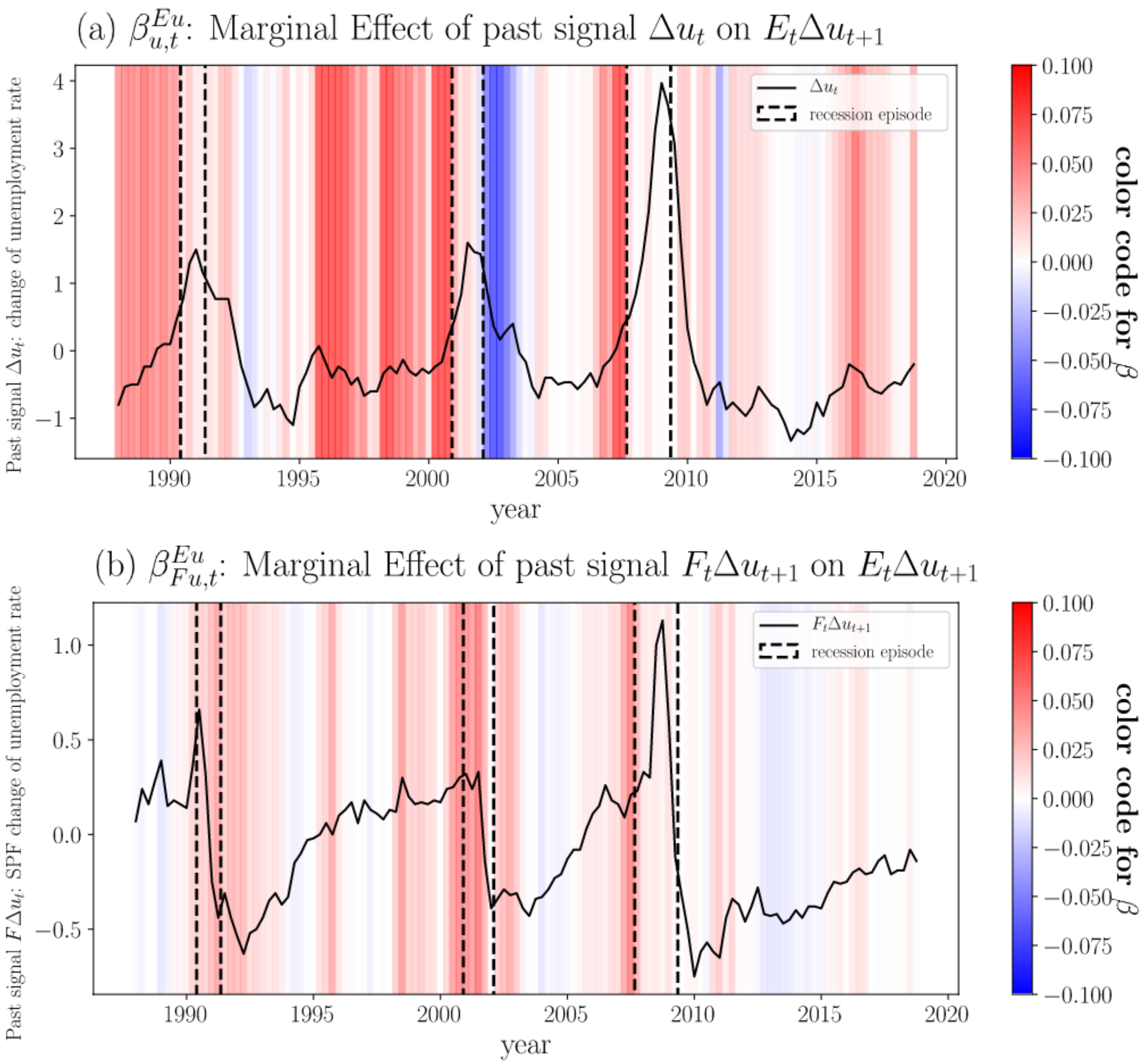


Figure 7:

To show the same attention shift pattern holds for all signals and expectations related to economic condition, I first plot the same heatmap for marginal effect of unemployment signals on expectation on unemployment change. This is Figure 7 below. It shows the same pattern holds as in Figure 3: in recession marginal effect of future signal is bigger and the opposite is true for past signal.

For marginal effects of cross-signals, for example, the impact of unemployment signal on economic condition expectation. These results are shown in Figure 8 below. It shows first unemployment signals generally have negative impact on expectation of economic condition. Furthermore, when looking at marginal effects of past signals, such an impact is again weak during recession periods whereas the marginal effects of future signals are again with bigger magnitudes during recessions.

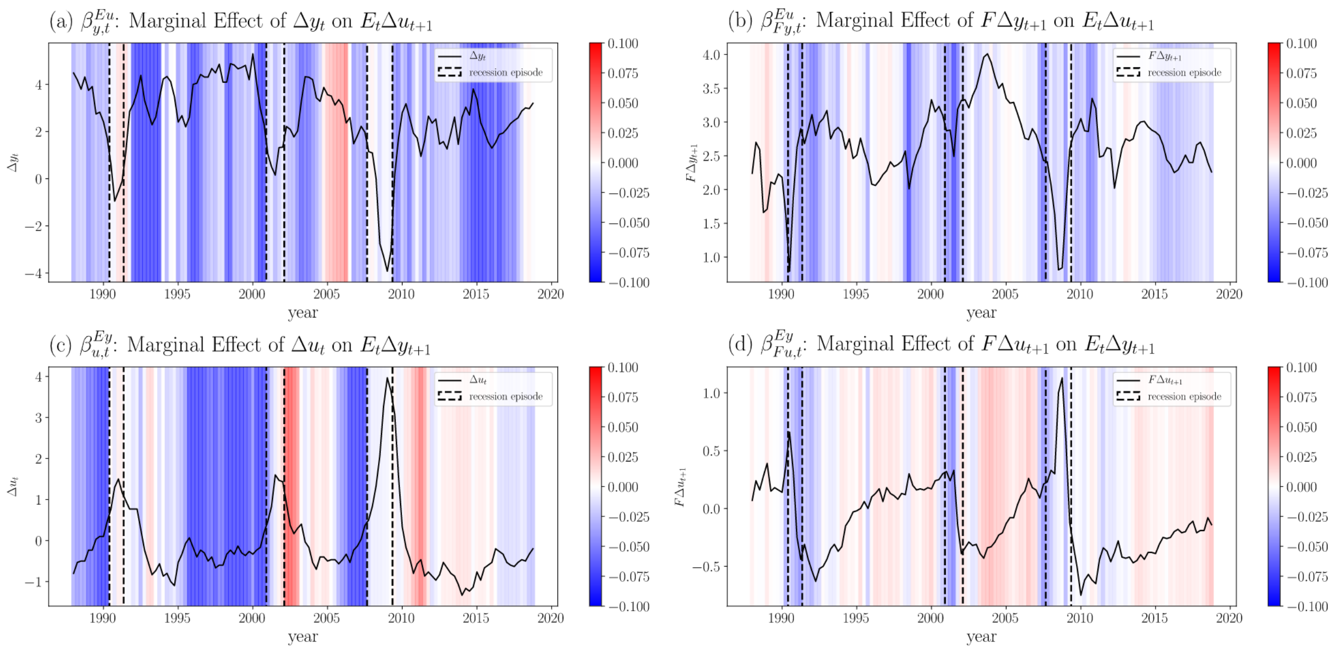


Figure 8:

However these attention shift during recession and ordinary period only holds significantly for expectations and signals related to indicators about economic conditions. Figure 9 plots the time-varying marginal effects for indicators on inflation and interest rate, there is no such attention shift at presence. The DML estimator also suggest the average marginal effects in recession and ordinary periods are not significantly different.

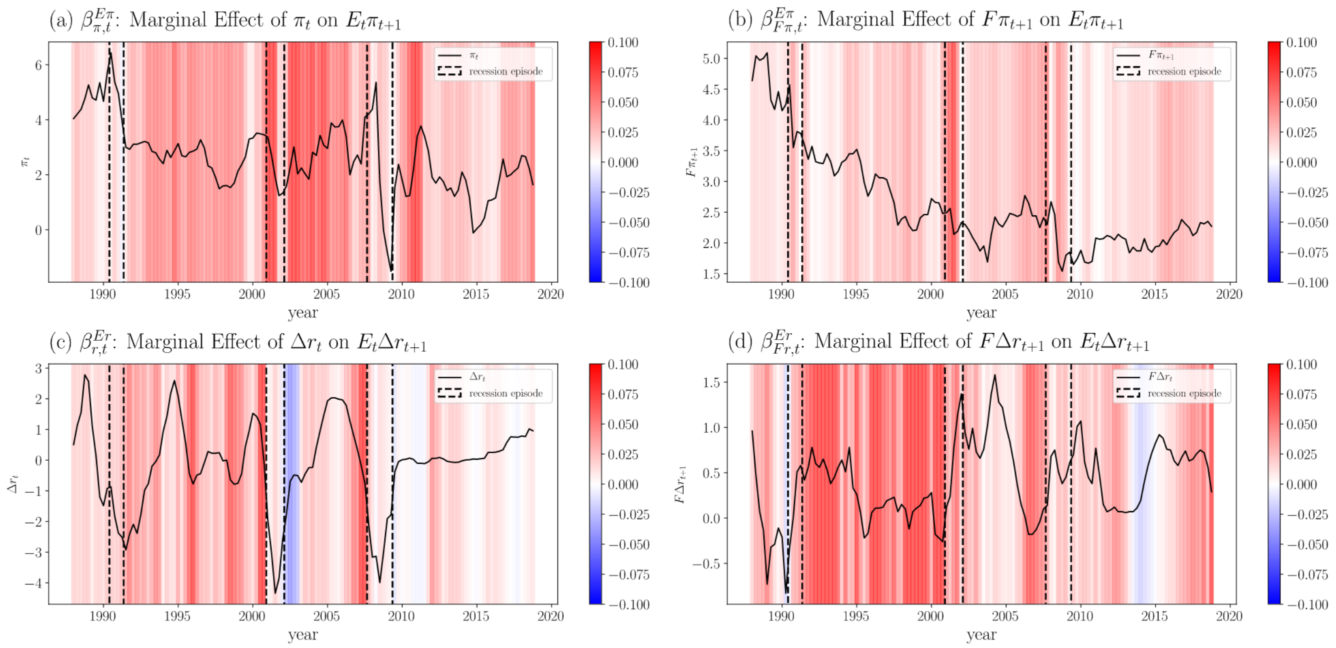


Figure 9:

D.2 Robustness of DML using NBER Recessions

Table 8: Average Marginal Effect of Past and Future Signals: NBER Recession

Expectation:		$E\Delta y_{t+1\|t}$			$E\Delta u_{t+1\|t}$		
	Signal	β_{bad} (std)	β_{ord} (std)	$\beta_{bad} = \beta_{ord}$ (p-val)	β_{bad} (std)	β_{ord} (std)	$\beta_{rec} = \beta_{ord}$ (p-val)
Future Signal	$F_t\Delta u_{t+1}$	**−0.047*** **	0.005	¡0.01	**0.033*** **	0.009***	¡0.01
		(0.006)	(0.002)		(0.004)	(0.002)	
	$F_t\Delta y_{t+1}$	**0.05*** **	0.02***	¡0.01	**−0.024*** **	−0.01***	¡0.01
		(0.007)	(0.003)		(0.003)	(0.001)	
Past Signal	Δu_t	−0.016*	−0.018***	0.86	0.012***	0.01***	0.74
		(0.008)	(0.003)		(0.005)	(0.002)	
	Δy_t	0.003	**0.015*** **	0.05	−0.004**	**−0.01*** **	0.04
		(0.004)	(0.002)		(0.002)	(0.001)	

* ***,**,*: Significance at 1%,5% and 10% level. β_{bad} is average marginal effect in bad periods defined by NBER recession dates, β_{ord} is average marginal effect in ordinary period. $\beta_{bad} = \beta_{ord}$ is test on equality between average marginal effects, its p-value is reported for each expectation-signal pair. Bold estimates denote the marginal effect with significantly bigger magnitude. Standard errors are adjusted for heteroskesticity and clustered within time.

Table 8 shows the DML estimates for marginal effects of past and future signals on real GDP growth and unemployment rate, during or out of recession. And the recession dates in use are those from NBER. Although "bad times" defined in **Section 4.2.2** are considered more plausible for reasons discussed before, using NBER recession dates won't qualitatively change the DML estimates much. Future signals still significantly have higher weights during bad periods and the weights on past signals are usually with bigger magnitude in ordinary period.

D.3 Decompose Time-varying ME with other signals

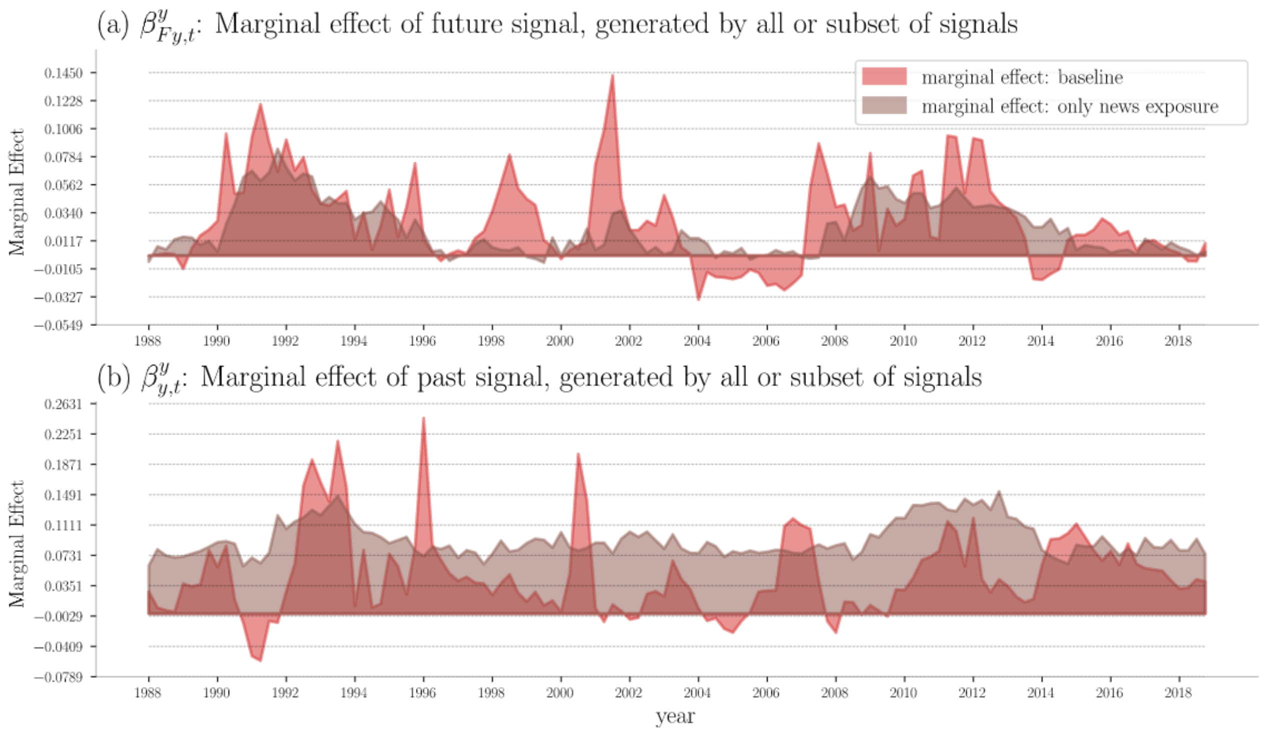


Figure 10: Time-varying marginal effect of past and future signal on real GDP growth. Top panel: marginal effect of future signal, $\beta^{Ey}_{Fy,t}$; bottom panel: marginal effect of future signal, $\beta^{Ey}_{y,t}$. The red curve: marginal effect created by estimated ASF with all signals. The blue curve: marginal effect created by ASF with only exposure of economic condition news.

Figure 10 presents how news exposure affects the marginal effects on future and past signals. It shows that news exposure only creates higher weights on future signals (from SPF) exactly when there is more news on economic status but not the weight change of past signals. According to Table 4, news exposure only accounts for 28% and 15% time-variation of weights on future and past signals, whereas economic conditions alone explain more than 50%. These suggest the explanation that attention-shift is majorly a result of more information available in recessions is unlikely to be true. On the other hand, economic condition signals without news exposure successfully recreate the key attention-shift pattern. This indicates economic condition signals explain a much bigger fraction of the time variation and are likely to be the main driving force for attention shift.

In Figure 11 I report the same exercise as in Figure 4 and 10 but with only signals on inflation, interest rate, and oil prices as input. The results show that information on prices alone cannot recreate the attention-shift pattern.

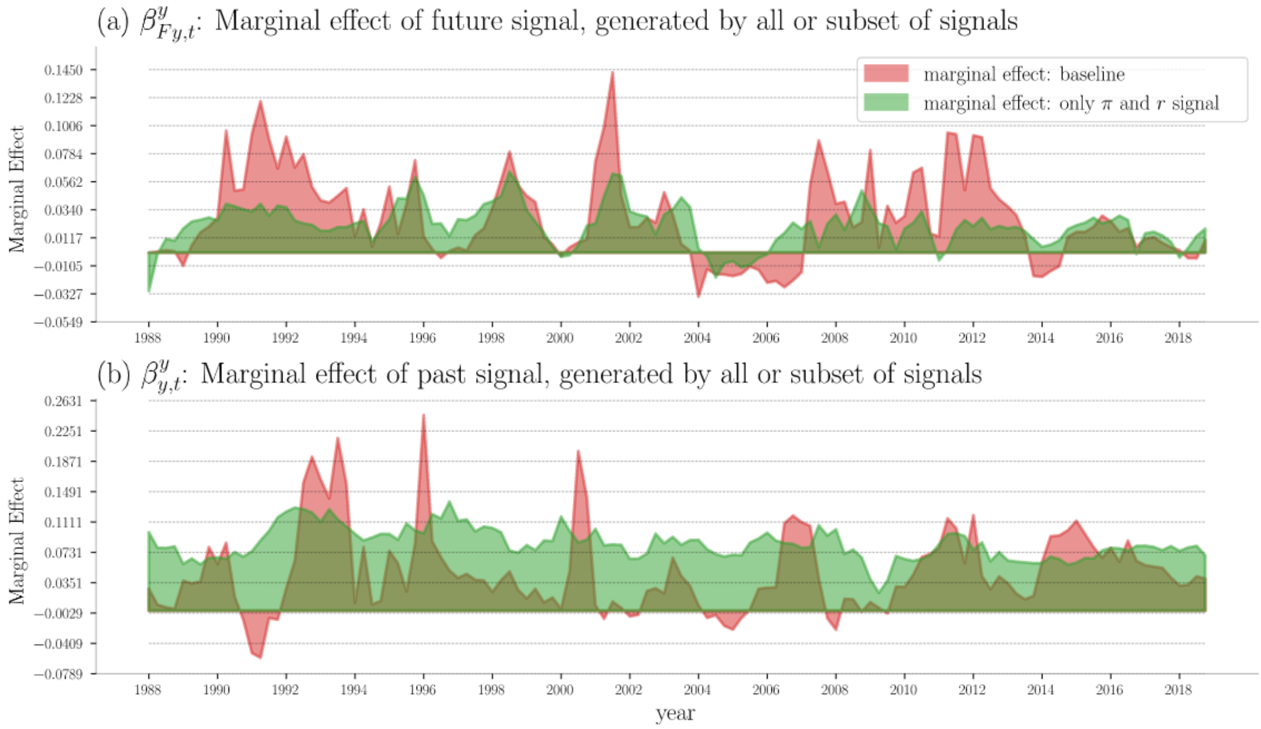


Figure 11: Time-varying marginal effect of past and future signal on real GDP growth. Top panel: marginal effect of future signal, $\beta^{Ey}_{Fy,t}$; bottom panel: marginal effect of future signal, $\beta^{Ey}_{y,t}$. The red curve: marginal effect created by estimated ASF with all signals. The blue curve: marginal effect created by ASF with only interest rate and inflation signals.

D.4 Variance Decomposition for Unemployment Expectation

In Table 9 I summarize the variance decomposition of time varying marginal effects of unemployment signal on unemployment expectations. It is consistent with what I find for expectation on economic condition. First the signals that explain most of the time-variation are those related to economic conditions. News exposure also explain a significant part of variation, especially for past signals. Finally these signals affect expectations through both accumulated states and covariates. Current signal usually plays a more important role in explaining the time-variation.

E Model Appdendix

Table 9: Variance Decomposition of Time-varying Marginal Effects: $E\Delta u$

Marginal Effect on Past Signal:		$\beta^{Eu}_{u,t}$				
Signal Type:		Economic Condition	Inflation	Interest rate	News	Total
	State $\theta_{i,t-1}$	28%	3%	6%	20%	57%
Channel:	Covariate $Z_{i,t}$	23%	2%	13%	5%	43%
	Total	52%	5%	18%	25%	
Marginal Effect on Future Signal:		$\beta^{Eu}_{Fu,t}$				
Signal Type:		Economic Condition	Inflation	Interest rate	News	Total
	State $\theta_{i,t-1}$	19%	6%	7%	4%	36%
Channel:	Covariate $Z_{i,t}$	36%	4%	9%	15%	64%
	Total	54%	10%	16%	19%	

E.1 Signals and Beliefs

At the beginning of time t, agent is endowed with some prior beliefs on states d_t and η_t, this reflects the latent states in empirical part. I denote the prior of foundamentals as:

$$\boldsymbol{X}_0 \equiv \begin{bmatrix} d_0 \\ \eta_0 \end{bmatrix} \sim N(\hat{\boldsymbol{X}}_0, \Sigma_0)$$

Where $\hat{\boldsymbol{X}}_0$ stands for prior mean of the states $\boldsymbol{X}_t$.[3]

The agent is Bayesian Learner and forms posterior beliefs using Kalman Filter. Agent updates his belief twice: first, he is exposed to a normal noisy signal z_0 about current state d_t. The variance of the noise is σ_z^2. The agent then updates her belief on $\boldsymbol{X}_t$. Because both prior and noise are normally distributed, the updated prior is also normal.

$$\boldsymbol{X}_{t|0} \equiv \begin{bmatrix} d_{t|0} \\ \eta_{t|0} \end{bmatrix} \sim N(\hat{\boldsymbol{X}}_{t|0}, \Sigma_{t|0})$$

I define $\boldsymbol{X}_{t|0}$ as conditional prior as it contains information about d_t. Specifically, its mean $\hat{\boldsymbol{X}}_{t|0}$ is a function of the unconditional prior mean and signal z_0, which contains information about d_t and noise. However, the agent has no control of the variance of this noise σ_z^2. It will

[3]In steady state one can think of the prior mean being at the long-run mean of each state, which is 0. When an agent observes a history of signals before time t, she may have a prior mean different from 0. This then can be thought of as a form of the "internal states" described in Section 4.2.4.

not be in the agent's choice set and will be treated as given when the agent solves the rational inattention problem later.

The second time the agent updates belief is after observing signal $\boldsymbol{Z}_t$. He forms a posterior belief about the fundamentals next period. As this is a two-period model, only belief on d_{t+1} is relevant. Again the agent forms belief using Bayes Rule:

$$\boldsymbol{X}_{t+1|t} \equiv \begin{bmatrix} d_{t+1|t} \\ \eta_{t+1|t} \end{bmatrix} \sim N(\hat{\boldsymbol{X}}_{t+1|t}, \Sigma_{t+1|t})$$

Where posterior mean $\hat{\boldsymbol{X}}_{t+1|t}$ and variance $\Sigma_{t+1|t}$ is defined as:

$$\hat{\boldsymbol{X}}_{t+1|t} \equiv \mathbb{E}[\boldsymbol{X}_{t+1}|\boldsymbol{Z}_t] = A\Big((I - KG)\hat{\boldsymbol{X}}_{t|0} + K\boldsymbol{Z_t}\Big) \tag{49}$$

$$\Sigma_{t+1|t} = A\Sigma_{t|0}A' - AKG\Sigma_{t|0}A' + \boldsymbol{Q} \tag{50}$$

And Kalman Gain is given by (51), where matrices A and G are given by exogenous parameters $\{\rho, \rho_\eta\}$ about the fundamentals.

$$K = \Sigma_{t|0}G'(G\Sigma_{t|0}G' + R)^{-1} \tag{51}$$

From (49)-(51), the choices of signal precision will affect both mean and variance of his posterior through the variance-covariance matrix on noise, R. Signals with lower variance are more accurate, and the agent will put higher weights on these signals. Each different choice of signal accuracy (represented by the variance-covariance matrix on noise, R) gives the agent a different information set. Given different information sets, the agent will form different posterior beliefs even if the signals realized are the same.

E.2 Two Special Information Set

At this point, it is worth describing several special information sets:

d_t Fully Observable: At time t, an agent has only perfect information about d_t and no information on η_t. This happens when $\sigma_{2,\xi} = 0$ and $\sigma_{1,\xi} \to \infty$. In this case agent will form adaptive expectation about return in the future: $E_t^A d_{t+1} = \rho d_t$.

Both fundamentals $\boldsymbol{X}_t$ Fully Observable: At time t, an agent has all the information about fundamentals at time t. Given the distribution of $\epsilon_{1,t+1}$, an agent with this information set can form a posterior belief on the distribution of d_{t+1} with mean being expressed as (23). This can be thought of as the Full Information Rational Expectation benchmark in this model as the forecasting error in this case will only be the unpredictable shock $\epsilon_{1,t+1}$.

An information set with an arbitrary variance-covariance matrix on noise, R, can be thought of as being in the middle of the two information sets described above. For each information set $\mathcal{I}_t$ given, the agent will solve her optimization problem (17) accordingly. Different information sets will then result in different choices, thus giving the agent different expected utility. In this sense, information has a value that can be evaluated with her expected utility. I will first illustrate the value of information in this model using different information sets described here.

E.3 State-dependent Value of Information

In this section, I explicitly compute the agent's expected utility conditional on different information sets given. I will illustrate that more information is valuable to agents as it increases their expected utility. Furthermore, the improvement of expected utility obtained by possessing more information depends on the current state of the economy, d_t.

I solve problem (18) under the two different information sets introduced before as well as the case with no extra information. Then I evaluate the agent's ex ante expected utility.

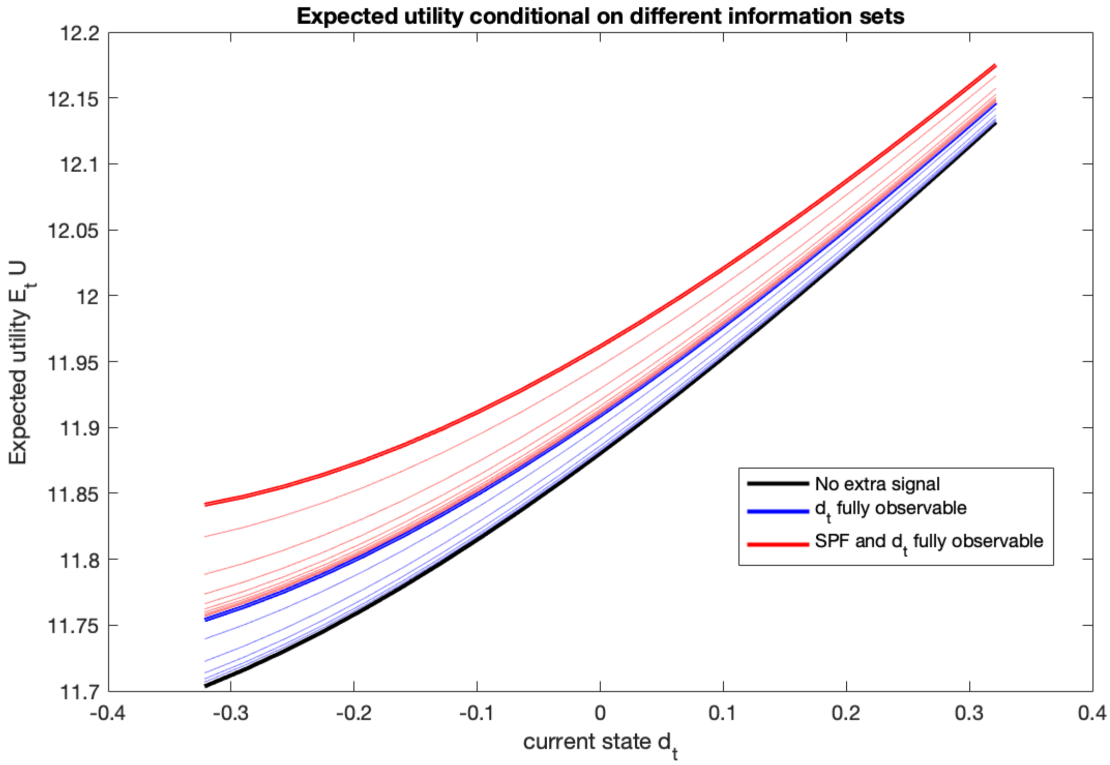
$$\mathbb{E}[u(e_t - s^*_{t+1}(\mathcal{I}_t)) + \beta u(r_{t+1}s^*_{t+1}(\mathcal{I}_t) + e_{t+1})|\mathcal{I}_0]$$

Recall the utility function takes the form $u(c_t) = c_t - bc_t^2$, and in the information set $\mathcal{I}_0$, it contains information about the current state d_t. The state-dependency seen later comes from the fact that the value of information changes as the mean of prior contained in $\mathcal{I}_0$ changes. The quadratic function form makes the point that the common mean-independent result of the rational inattention model is not due to linear quadratic preference per se, rather it's because of the quadratic approximation for the entire problem. However, the results are not restricted to such a utility function form. Recall that given different information set $\mathcal{I}_t$, the first order condition for problem (18) takes the form:

$$s^*_{t+1}(\mathcal{I}_t) = \frac{-1 + 2be_t + (\beta - 2b\beta e_{t+1})\mathbb{E}[r_{t+1}|\mathcal{I}_t]}{2b(1 + \beta\mathbb{E}[r^2_{t+1}|\mathcal{I}_t])} \tag{52}$$

For illustration purposes, I solve the model numerically using the following parametrization: $b = 1/40$, $\beta = 0.95$, $e_t = 10$ and $e_{t+1} = 5$. For the fundamentals I consider $\rho = 0.2$, $\rho_d = 0.9$, $\sigma_{1,\epsilon} = \sigma_{2,\epsilon} = 0.15$. The prior beliefs on states d_t and η_t are assumed to be mean zero with the stationary variance-covariance matrix obtained from the recursive Kalman Filter. The standard deviation of noise on passive signal is $\sigma_z = 0.22$. In Figure 12 I plot the ex ante expected utility conditional on various information sets, as functions of realized d_t. The thick black curve is expected utility when there's no more information other than the initial passive signal on d_t available to the agent. The thick blue curve is expected utility when d_t is fully observable and the thick red curve is when both SPF and d_t are fully observable (the FIRE benchmark)[4]. The curves between the thick lines depict the increase in expected utilities as the precision of the signal increases (or the variance of noise decreases).

Figure 12: Expected Utility under Different Information Set



Black thick line: Expected utility when no more information other than initial passive signal on d_t; blue thick line: expected utility when d_t becomes fully observable; red thick line: when both SPF and d_t fully observable. Blue thin lines are expected utilities when there are noise attached to extra signal on d_t, the more accurate the signal, the closer it gets to d_t fully observable case. Red thin lines are expected utilities when noise attached to signal on SPF, and d_t is fully observable. The more accurate the signal, the closer it gets to full information case.

There are two key messages from Figure 12. First, more information improves the agent's expected utility progressively: with a more accurate signal on d_t, the agent resolves the uncertainty about the current state and his utility increases at any given d_t from the black line to the blue line; and it continues to increase as the signal on SPF becomes more accurate, from blue curve to red curve. This is a typical result of informational models.

Secondly and more importantly, the value of information is decreasing in realized state d_t. This can be seen from the differences between expected utilities with different information sets. When realized state d_t is low and negative, getting the same amount of information will increase the agent's expected utility by more than the case when d_t is high. In other

[4]With the specific law of motion assumed in (19) - (21) together with definition of SPF (23), the case with only SPF fully observable will coincide with the FIRE case.

words, information is more valuable when the economic status is bad. This is a result different from that of standard rational inattention literature. The reason for such a difference is the non-linearity in the optimal saving/investment function.

The difference between expected utility comes from differences of optimal investment (52). The fact that optimal saving is a non-linear function of both posterior mean and posterior variance of state d_{t+1} makes the expected utility mean-dependent. To see this, we can utilize the assumption of the quadratic utility function, and re-write the expected utility as the following form:

$$\begin{aligned}
\mathbb{E}[U(s^*_{t+1}(\mathcal{I}_t))] &= \mathbb{E}[(e_t - s_{t+1}) - b(e_t - s_{t+1})^2 + \beta(e_{t+1} + r_{t+1}s_{t+1}) - \beta b(e_{t+1} + r_{t+1}s_{t+1})^2] \\
&= \mathbb{E}[-\underbrace{b(1+\beta r^2_{t+1})}_{\equiv\chi} s^2_{t+1} + (2be_t - 1 + \beta r_{t+1} - 2\beta b e_{t+1} r_{t+1})s_{t+1} + (e_t - be_t^2 + \beta e_{t+1} - \beta b e^2_{t+1})] \\
&= \mathbb{E}[-\chi(s^2_{t+1} - \underbrace{\frac{2be_t - 1 + \beta r_{t+1} - 2\beta b e_{t+1} r_{t+1}}{\chi}}_{\equiv 2\bar{s}_{t+1}} s_{t+1} + \frac{(2be_t - 1 + \beta r_{t+1} - 2\beta b e_{t+1} r_{t+1})^2}{4\chi^2}) \\
&+ \frac{(2be_t - 1 + \beta r_{t+1} - 2\beta b e_{t+1} r_{t+1})^2}{4\chi} + (e_t - be_t^2 + \beta e_{t+1} - \beta b e^2_{t+1})] \\
&= -\mathbb{E}[\chi(s_{t+1} - \bar{s}_{t+1})^2] + \underbrace{\mathbb{E}[\frac{(2be_t - 1 + \beta r_{t+1} - 2\beta b e_{t+1} r_{t+1})^2}{4\chi} + (e_t - be_t^2 + \beta e_{t+1} - \beta b e^2_{t+1})]}_{\equiv M}
\end{aligned}$$

Note in the above derivation all the expectations are conditional on initial information set $\mathcal{I}_0$. M has nothing to do with information set $\mathcal{I}_t$, thus the evaluating the expected utility under choice of $\mathcal{I}_t$ is equivalent to evaluating the quadratic loss term $\mathbb{E}[\chi(s^*_{t+1}(\mathcal{I}_t) - \bar{s}_{t+1})^2]$. This is a standard result from literature of Rational Inattention with linear quadratic preference. The key difference here is s^*_{t+1} is non-linear in fundamentals.[5] We can then write ante expected utility as a form of "quadratic loss":

$$\mathbb{E}[U(s^*_{t+1}(\mathcal{I}_t))|\mathcal{I}_0] = -\mathbb{E}[\chi(s^*_{t+1}(\mathcal{I}_t) - \bar{s}_{t+1})^2|\mathcal{I}_0] + M \tag{53}$$

The variable $\bar{s}_{t+1}$ is given by (54). It stands for the optimal investment under perfect foresight when the agent observes d_{t+1} perfectly.

$$\bar{s}_{t+1} = \frac{-1 + 2be_t + \beta r_{t+1} - 2b\beta r_{t+1} e_{t+1}}{2b(1+\beta r^2_{t+1})} \tag{54}$$

The transformed utility function (53) is usually referred to as a quadratic loss function in rational inattention models, intuitively agent will seek to minimize the expected loss between optimal choice under limited information set $\mathcal{I}_t$ and optimal choice under Full Information Rational Expectation.[6] From (53) it is obvious that if the optimal choice of s is linear in state

[5]In standard rational inattention models, the action will be linear in fundamentals thus optimal choice of signal will not depend on prior mean of fundamentals. For example, see Maćkowiak *et al.* (2018) or Kamdar (2019).

[6]It is worth noting that M is not involved in choosing the optimal information structure $\mathcal{I}_t$ as it is only related to the actual distribution of r_{t+1}.

r_{t+1}, the expected utility only depends on the posterior variance of r_{t+1} given information set $\mathcal{I}_t$. It is not related to the posterior mean of states or realized state at time t.

Using the transformed expected utility (53), I can explore reasons for the value of information decreasing in d_t. To see this, consider the cases with or without full information from SPF. Conditional on the realization of a specific d_t, without information from SPF agent faces uncertainty from both η_t and $\epsilon_{1,t+1}$ being unobservable. With information from SPF uncertainty from η_t is resolved. Because in both cases agents have no information on $\epsilon_{1,t+1}$, the utility improvement comes solely from knowledge on η_t. For simplicity, I consider an extreme case when $\epsilon_{1,t+1} = 0$. Then $\bar{s}_{t+1}$ can be seen as the optimal saving choice when SPF is available. The utility loss of the agent not having information from SPF can then be evaluated by differences between optimal savings with or without SPF observable, weighted by the agent's subjective belief.

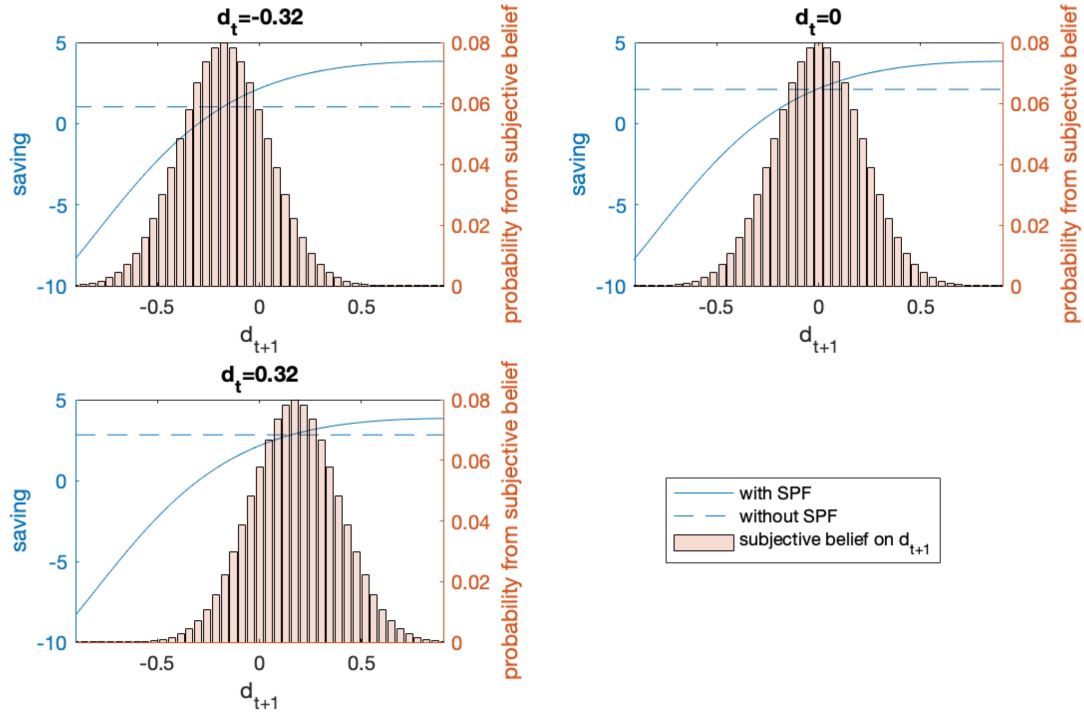
In Figure 13 I depict the optimal saving choices at three realized values of d_t: when the current state is bad ($d_t = -0.32$), neutral ($d_t = 0$) and good ($d_t = 0.32$). In each case, I plot the optimal saving choice as a function of future state d_{t+1}. The dotted line is the optimal saving that the agent chooses when he only observes the initial signal on d_t. It is a flat line because the agent's choice does not depend on d_{t+1} (or realization of η_t) when SPF is not observable. The solid line is the optimal saving choice when SPF is observable to the agent. This line is a function of d_{t+1} because under the assumption $\epsilon_{1,t+1} = 0$, when SPF is observable then η_t and d_{t+1} are fully observed. An important feature is then this function is increasing and concave in d_{t+1}. This is because the higher the return d_{t+1} is, the more agent wants to save. The concavity comes from the fact that the substitution effect becomes weaker as the return on asset increases and is finally dominated by the income effect.[7]

Now for agents without information from SPF, the solid line is not feasible. For a given realized d_t, the agent will evaluate her utility loss of not having information on η_t following (53). This is done by measuring the distance between optimal saving choices with and without information from SPF and computing the expected value of (the square of) this distance using their posterior belief on d_{t+1} (η_t). In Figure 13 this belief is shown with a bar plot. When realized d_t is higher, the belief of distribution on d_{t+1} is centered at a higher mean. Because of the non-linearity of the optimal saving choice, the average distance between saving choices with and without information from SPF is higher when d_t is low. This gives rise to the fact that value of information from SPF is decreasing in d_t.

With the simple structure presented above, I show the key pattern my model generates: the value of information decreases in the state of the economy. The agent is willing to pay higher costs to acquire information as the state of the economy gets worse. This gives the key mechanism to create the time-varying marginal effect and non-linearity I documented

[7]Interestingly, if one would instead assume a riskless asset with a risky endowment in $t + 1$, the optimal saving curve under full information will be linear and the value of information won't depend on the current state anymore.

Figure 13: Optimal Saving under Full Information and Limited Information



Solid line: optimal saving choice under full information: when both d_t and SPF are fully observable. Dash line: optimal policy when SPF signal is not available. Bar plot: agent's subjective belief on future state d_{t+1}, when SPF is not observable. Top left panel is when current state is very bad ($d_t = -0.32$), top right panel is when $d_t = 0$ and bottom left panel is when current state very good, $d_t = 0.32$.

with RNN. Because when agents can choose the precision of signals (thus information set) optimally, they will make different choices during bad and ordinary times and this will result in different weights on these signals.

E.4 Derivation of Information Cost

In this subsection, I derive the information cost measured by entropy in (25) following Mackowiak and Wiederholt (2009). Recall the state-space representation of fundamentals are:

$$\boldsymbol{X}_{t+1} = A\boldsymbol{X}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{Q})$$

The initial noisy signal z_0 and chosen signals $\boldsymbol{Z}_t$ are given by:

$$z_0 = d_t + \xi_0 = G_0\boldsymbol{X}_t + \xi_0, \quad G_0 = [1 \quad 0]$$

$$\boldsymbol{Z}_t = G\boldsymbol{X}_t + \boldsymbol{\xi}_t, \quad \boldsymbol{\xi}_t \sim N(\boldsymbol{0}, R)$$

First notice all the random variables $\boldsymbol{X}_t$, z_0, $\boldsymbol{Z}_t$ are normally distributed. The information set $\mathcal{I}_0$ only contains a noisy Gaussian signal z_0, the entropy of $\boldsymbol{X}_{t+1}$ given $\mathcal{I}_0$ is then:

$$\mathcal{H}(\boldsymbol{X}_{t+1}|\mathcal{I}_0) = \mathcal{H}(\boldsymbol{X}_{t+1}|z_0) = \frac{1}{2}log_2[(2\pi e)^2 det\Sigma_{t+1|0}] \tag{55}$$

Where $\Sigma_{t+1|0}$ denotes the variance-covariance matrix of $\boldsymbol{X}_{t+1}$ given z_0. The prior variance covariance matrix of $\boldsymbol{X}_t$ is denoted as Σ_0, then the conditional variance-covariance matrix $\Sigma_{t+1|0}$ is given by:

$$\Sigma_{t+1|0} = A\Sigma_{t|0}A' + \boldsymbol{Q} \tag{56}$$

Where:

$$\Sigma_{t|0} = \Sigma_0 - \Sigma_0 G_0'(G_0\Sigma_0 G_0' + \sigma_z^2)^{-1}G_0\Sigma_0 \tag{57}$$

It is obvious $\Sigma_{t|0}$ by construction the posterior variance-covariance matrix for hidden states $\boldsymbol{X}_t$ after observing z_0 derived from Kalman Filter.

Then recall $\mathcal{I}_t = \mathcal{I}_0 \cup \{\boldsymbol{Z}_t\}$, similar as above we have:

$$\mathcal{H}(\boldsymbol{X}_{t+1}|\mathcal{I}_t) = \mathcal{H}(\boldsymbol{X}_{t+1}|z_0, \boldsymbol{Z}_t) = \frac{1}{2}log_2[(2\pi e)^2 det\Sigma_{t+1|t}] \tag{58}$$

Where:

$$\Sigma_{t+1|t} = A\Sigma_{t|0}A' - A\Sigma_{t|0}G'(G\Sigma_{t|0}G' + R)^{-1}G\Sigma_{t|0}A' + \boldsymbol{Q} \tag{59}$$

Again by construction, the $\Sigma_{t+1|t}$ is the posterior variance-covariance matrix for $\boldsymbol{X}_{t+1}$ after observing $\{z_0, \boldsymbol{Z}_t\}$ derived from Kalman Filter. Moreover, it is obvious the uncertainty after observing $\boldsymbol{Z}_t$ is reduced compared to the uncertainty after only observing z_0.

Now information cost is obtained by measuring uncertainty reduction induced by extra information, using (58) and (55) we have the information cost in (25):

$$\mathcal{H}(\boldsymbol{X}_{t+1}|\mathcal{I}_0) - \mathcal{H}(\boldsymbol{X}_{t+1}|\mathcal{I}_t) = \frac{1}{2}log_2(\frac{det\Sigma_{t+1|0}}{det\Sigma_{t+1|t}})$$

E.5 Derivation of $\mathbb{E}[r_{t+1}|\mathcal{I}_t]$ and $\mathbb{E}[r_{t+1}^2|\mathcal{I}_t]$

From (49):

$$\begin{pmatrix} \mathbb{E}[d_{t+1}|\mathcal{I}_t] \\ \mathbb{E}[\eta_{t+1}|\mathcal{I}_t] \end{pmatrix} \equiv \hat{\boldsymbol{X}}_{t+1|t} = A\Big((I - KG)\hat{\boldsymbol{X}}_{t|0} + K\boldsymbol{Z_t}\Big)$$

Where $\hat{\boldsymbol{X}}_{t|0} = \mathbb{E}[\boldsymbol{X}_t|\mathcal{I}_0]$ is the mean of belief on $\boldsymbol{X}_t$ after observing passive signal z_0. The prior before observing z_0 is denoted as $\boldsymbol{X}_0 \sim N(\hat{\boldsymbol{X}}_0, \Sigma_0)$ from Section E.1. Now denote the Kalman Gain for observing z_0 as K_0, we can write:

$$\hat{\boldsymbol{X}}_{t|0} = (I - K_0G_0)\hat{\boldsymbol{X}}_0 + K_0 z_0 \tag{60}$$

Where $G_0 = \iota = [10]$ as defined in Appendix E.4 and K_0 is given by:

$$K_0 = \Sigma_0 G_0'(G_0 \Sigma_0 G_0' + \sigma_z^2)^{-1} \tag{61}$$

Combine (49), (60) and (61) we have:

$$\begin{aligned}\mathbb{E}[r_{t+1}|\mathcal{I}_t] &= 1 + \iota A\Big((I - KG)\big((I - K_0 G_0)\hat{\boldsymbol{X}}_0 + K_0 z_0\big) + KZ_t\Big) \\ &= 1 + \iota A\Big((I - KG)\big((I - \Sigma_0 G_0'(G_0\Sigma_0 G_0' + \sigma_z^2)^{-1} G_0)\hat{\boldsymbol{X}}_0 + \Sigma_0 G_0'(G_0 \Sigma_0 G_0' + \sigma_z^2)^{-1} z_0\big) + KZ_t\Big)\end{aligned} \tag{62}$$

Before I show the derivation of $\mathbb{E}[r_{t+1}^2|\mathcal{I}_t]$, it's useful to consider what is $Var(d_{t+1}|\mathcal{I}_t)$. It is the first element of posterior variance covariance matrix $\Sigma_{t+1|t}$, which is given by (59). So $Var(d_{t+1}|\mathcal{I}_t)$ can be written as:

$$\begin{aligned}Var(d_{t+1}|\mathcal{I}_t) &= \iota \Sigma_{t+1|t} \iota' \\ &= \iota\big(A\Sigma_{t|0}A' - A\Sigma_{t|0}G'(G\Sigma_{t|0}G' + R)^{-1} G\Sigma_{t|0}A' + \boldsymbol{Q}\big)\iota'\end{aligned} \tag{63}$$

Now we can derive $\mathbb{E}[r_{t+1}^2|\mathcal{I}_t]$:

$$\begin{aligned}\mathbb{E}[r_{t+1}^2|\mathcal{I}_t] &= Var(r_{t+1}|\mathcal{I}_t) + (\mathbb{E}[r_{t+1}|\mathcal{I}_t])^2 \\ &= Var(d_{t+1}|\mathcal{I}_t) + (\mathbb{E}[r_{t+1}|\mathcal{I}_t])^2 && (64) \\ &= \iota\big(A\Sigma_{t|0}A' - A\Sigma_{t|0}G'(G\Sigma_{t|0}G' + R)^{-1}G\Sigma_{t|0}A' + \boldsymbol{Q}\big)\iota' + (\mathbb{E}[r_{t+1}|\mathcal{I}_t])^2 && (65)\end{aligned}$$

In the above equation, $\Sigma_{t|0}$ is given by (57), which contains σ_z^2 and prior variance Σ_0. $\mathbb{E}[r_{t+1}|\mathcal{I}_t]$ is given by (62), which depends on prior mean $\hat{\boldsymbol{X}}_0$, precision (variance) of the signal R and passive signal z_0. From (62) and (65), it is clear that the optimal saving choice is a non-linear function of all these variables related to the information friction.

Now to see how the *ex-post* weights on signal Z_t depend on variances of signals R, denote the weight on SPF signal as w_{spf} and weight on signal about current state as w_t, from (62) we have:

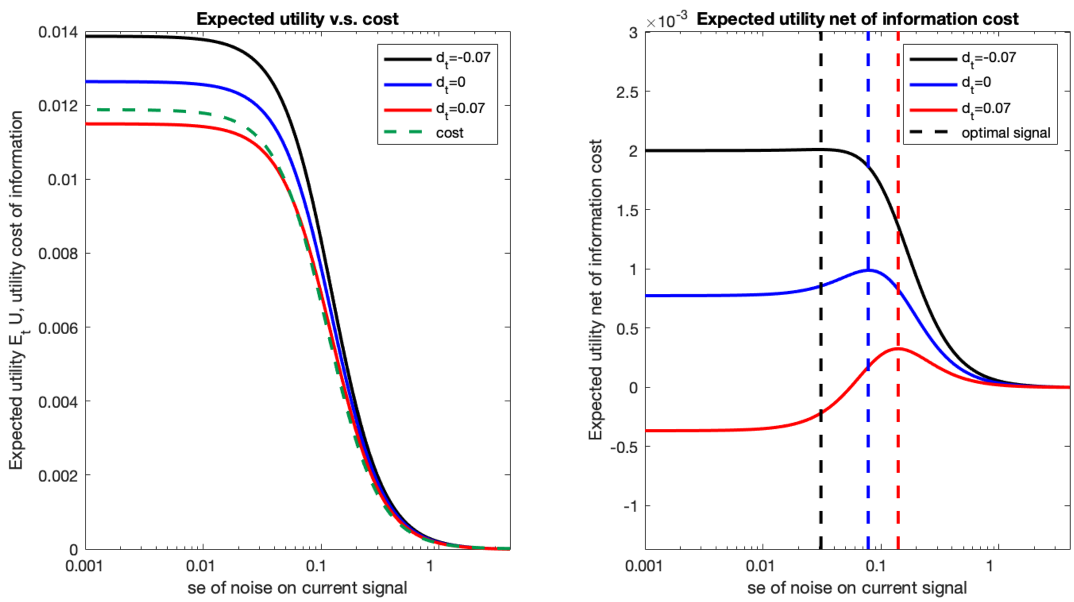
$$\begin{aligned}\begin{pmatrix} w_{spf} \\ w_t \end{pmatrix} &= (\iota' \quad \iota')AK \\ &= (\iota' \quad \iota')A\Sigma_{t|0}G'(G\Sigma_{t|0}G' + R)^{-1}\end{aligned} \tag{66}$$

From we see first for given Σ_{t+0}, G and A, a lower variance of noise on signal (contained in R) leads to higher weights put on corresponding signal. Moreover, as $\Sigma_{t|0}$ is affected by σ_z^2, the weights on signals also change with σ_z^2. This will be verified in Section E.7.1.

E.6 State-dependent Optimal Signals

Now turn to the rational inattention problem (26)-(28). I first show that the trade-off between the benefit and cost of acquiring information changes with the current state d_t. In Figure 14, I present the expected utility, information cost and the objective function in (26) when the current state d_t is negative, at mean zero and positive. The left panel describes how the expected utility changes as the standard error of the current signal ($\sigma_{2,\xi}$) changes for the three cases of d_t. Because at a different level of d_t, the expected utility for the same signal precision will be different, I normalize it by the utility at $\sigma_{2,\xi} = \infty$, which corresponds to the case when the agent acquires no extra signal on the current state. It is obvious in all three cases of d_t, the higher(lower) the precision(standard error) of the signal, the higher the expected utility comparing to the no information case. I then present the information cost for all three cases and show that the information costs are the same across different levels of d_t. This is because, in (30), the passive signal z_0 contains information about the current state d_t thus making the expected utility depending on it. Whereas in the information cost (32), the evaluation of posterior variance is mean-independent, which means only the variance of the passive signal matters in accessing the information cost so that the cost will not change as d_t changes. The key message from the left panel is that both the cost and the benefit of information increase with the precision of the current signal. Meanwhile for the same level of signal precision, the higher the current state d_t, the lower the benefit from that signal.

Figure 14: Information Benefit, Information Cost and Households' Objective: Function of Current Signal

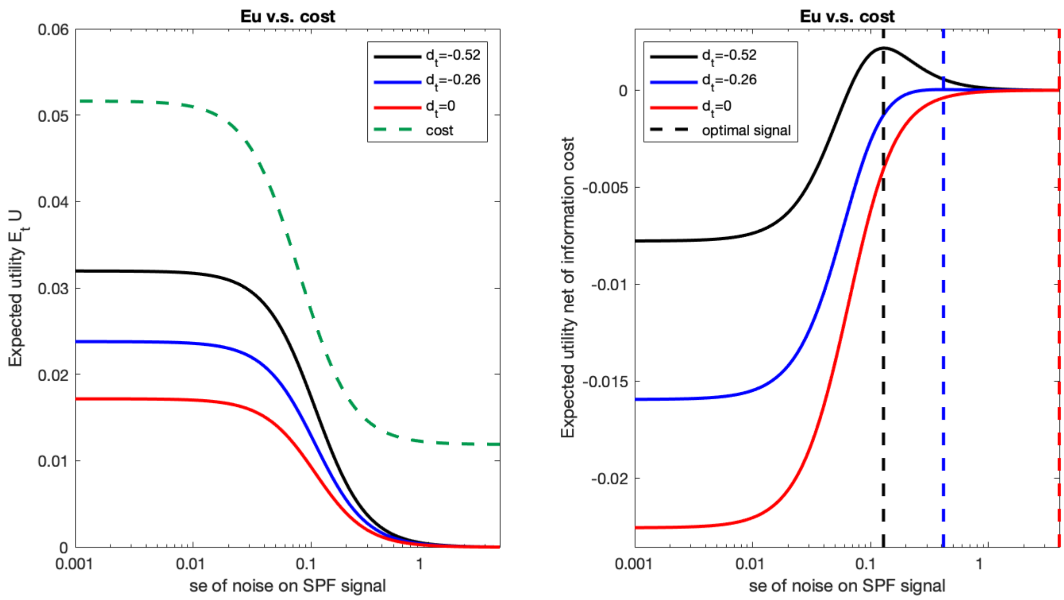


Left panel: the information benefit is evaluated by expected utility and plotted with solid lines, information cost is evaluated by entropy cost and plotted with dashed line. Right panel: objective function is obtained by information benefit minus cost and plotted with solid lines. The figure considers three different cases: current state is high with $d_t = 0.07$, current state is at its mean $d_t = 0$ and current state is low at $d_t = -0.07$. Horizontal axis is standard error of noise on *current signal*, higher s.e. leads to lower weight. Vertical dashed line in the right panel labels optimal s.e. for *current signal* in three scenarios.

The agent's objective function considers both the cost and benefit of acquiring information. The right panel of Figure 14 then presents the objective function under the same three cases of d_t. A utility-maximizing agent will choose the current signal with a standard error that maximizes her objective function. These choices under different states are represented by the dashed line. The right panel then shows that when the current state is worse, the agent will choose a higher precision and a lower standard error for the current signal.

These patterns hold true for precision on signals about SPF as well. In Figure 15, I again show a similar graph as in Figure 14 but for signals on SPF. The major difference between this figure and Figure 14 is that expected utility is computed assuming $\sigma_{2,\xi} = 0.001$, which means the agent chose a very precise current signal.[8] All the objects plotted in Figure 15 are then functions of the precision on SPF signal, $\sigma_{1,\xi}$. Similar to that in Figure 14, we see that both the benefit and the cost increase when the agent acquires more information on SPF. Meanwhile, the optimal precision of the SPF signal decreases with the current state d_t.

Figure 15: Information Benefit, Information Cost and Households' Objective: Function of SPF Signal



Left panel: the information benefit is evaluated by expected utility and plotted with solid lines, information cost is evaluated by entropy cost and plotted with dashed line. Right panel: objective function is obtained by information benefit minus cost and plotted with solid lines. The figure considers three different cases: current state $d_t = 0, -0.26$ and -0.52. For $d_t > 0$ the agent will always choose precision that leads to weight zero because here I plot all the objects under $\sigma_{2,\xi} = 0.001$. Horizontal axis is standard error of noise on *future (SPF) signal*, higher s.e. leads to lower weight. Vertical dashed line in the right panel labels optimal s.e. for *future (SPF) signal* in three scenarios.

[8] However, changing the level of $\sigma_{2,\xi}$, in this case, will not change the results qualitatively.

E.7 Comparative Statistics for Rational Inattention Model

I now examine the impacts of changing model parameters σ_z and prior mean $\hat{\boldsymbol{X}}_0$, on the optimal signal choices.

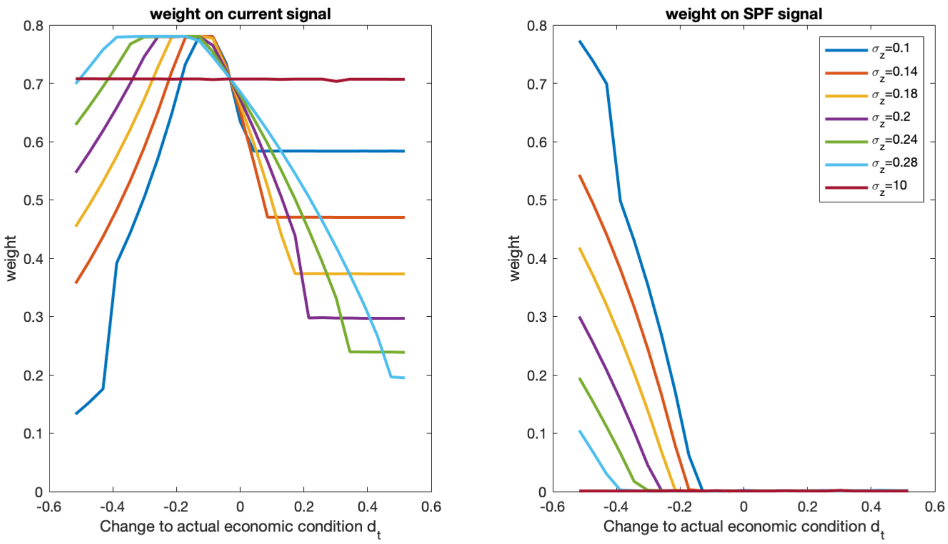
E.7.1 Impact of Passive Signal Precision σ_z

An important parameter in the information friction proposed in this chapter is the precision of the passive signal that the agent is exposed to. This passive signal z_0 will contain information about the current state d_t, thus making the agent's optimal choice of precision dependent on this state. The precision of this signal then determines the prior variances that the agent considers to evaluate the benefit and cost for more information. Through this channel, it will also affect the agent's optimal choices on signal precisions.

In Figure 16 I show again the optimal weights on current and SPF signals as functions of current state d_t, with different values of σ_z. The left panel shows the results for weights on current signals. First, notice higher precision on the passive signal (lower σ_z) leads to higher weight on the current signal to start with. This shows up in the figure as a higher weight in the flat area before the agent puts excess weight on the current signal. Intuitively, this means the agent already has a better understanding of the current d_t before choosing the extra signals on the current state and SPF. This leads to the fact that in the right panel, the agent with the lowest σ_z will start to pay attention to the SPF signal at a relatively higher state, because the information cost for choosing that precision level is relatively lower to her. An extreme example will be that when $\sigma_z = 0$, which implies that the agent has perfect information on d_t. In this case, we will see her only choosing an extra signal on SPF starting from a relatively high value of d_t.

Another interesting aspect in Figure 16 is that when the quality of the passive signal is very low so that σ_z is quite high, the agent's optimal choices of signal precisions will not depend on the current state d_t anymore. This is because the passive signal z_0 contains almost no information about d_t before the agent chooses her information set. As a result, the agent will not be able to choose different precisions according to the realization of d_t. This result is also shown in Figure 16 as the case for $\sigma_z = 0$. Moreover, in this case, the agent will not necessarily choose a very noisy signal about the current state. The optimal precisions will depend on the prior mean of the agent, which is $\hat{\boldsymbol{X}}_0 = \mathbf{0}$ as in the baseline results. Such a pattern then has an important implication: the weights on signals will not only depend on the realized current state of the economy d_t, it will also depend on the prior mean that the agent carried on across time. I will illustrate how the optimal weights change with the prior mean in the next subsection.

Figure 16: Weights on Signals: Noise on Passive Signal σ_z



Left panel: model implied weights on *current signal* as function of actual economic condition d_t. Right panel: model implied weights on *SPF signal*. Each set of weights corresponds to a different standard error of noise in the passive signal σ_z. The baseline results come from $\sigma_z = 0.18$.

E.7.2 Impact of "Internal State": Prior Mean

Intuitively, when an agent chooses the information set she uses to form expectations, her ex-ante belief about the future state should matter. This can be seen directly from (30): when the agent thinks about future state d_{t+1} before she chooses information set that will generate $\boldsymbol{Z}_t$, her effective prior mean should be:

$$\hat{\boldsymbol{X}}_{t|0} = (I - K_0 G_0)\hat{\boldsymbol{X}}_0 + K_0 z_0 \tag{67}$$

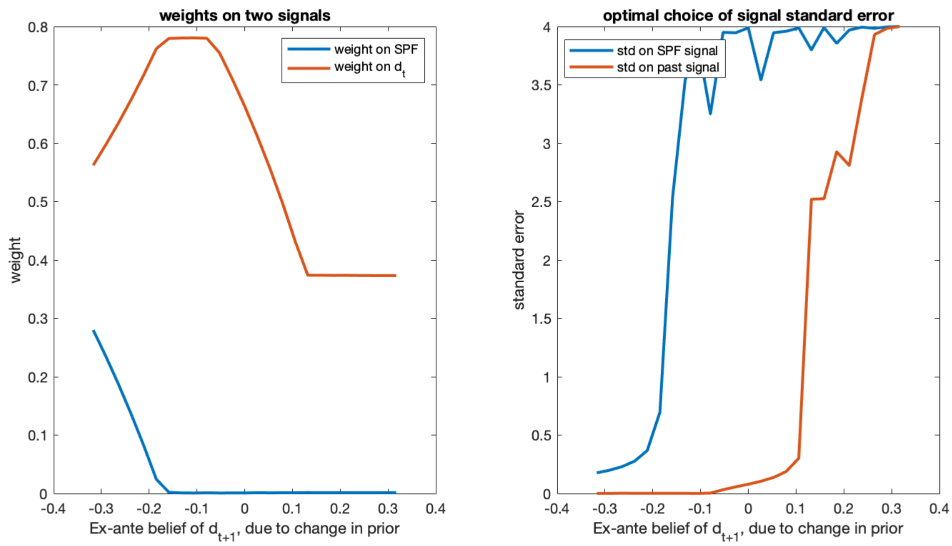
From previous sections, I have shown that the current state of economy d_t will affect her choice of optimal precision through the passive signal z_0. For the same reason, the optimal precision on signals should depend on the prior mean $\hat{\boldsymbol{X}}_0$ as well.

The illustration of the impact of prior mean involves several parts. First I want the change of optimal precision to come solely from the differences of $\hat{\boldsymbol{X}}_0$, so I will keep z_0 at a fixed value. Secondly, the reason why the prior mean will affect the optimal precision choice is that the agent will use the information set $\{\mathcal{I}_0\} = \{\boldsymbol{X}_{t|0}\}$ to evaluate her expected utility. The prior mean vector $\hat{\boldsymbol{X}}_0$ will affect this information set thus affecting the agent's expected benefit for any precision level. As discussed in section E.3, when the prior makes the agent believe on average the future state will be worse, she will choose a signal will higher precision. A straightforward way to illustrate this point is to depict the optimal weights and standard errors of the signals as functions of the implied posterior mean on d_{t+1} using the ex-ante information set $\mathcal{I}_0$. For simplicity, I call this "ex-ante belief on d_{t+1}", defined as:

$$\mathbb{E}[d_{t+1}|\mathcal{I}_0] = \iota A\Big((I - K_0 G_0)\hat{\boldsymbol{X}}_0 + K_0 z_0\Big) \tag{68}$$

In Figure 17, I show the optimal weights and standard errors as function of $\mathbb{E}[d_{t+1}|\mathcal{I}_0]$, while fixing $z_0 = 0$ and $d_t = 0$. This means the variation of the ex-ante belief comes solely from the differences in $\hat{\boldsymbol{X}}_0$.

Figure 17: Weights and Standard Error of Signals: Functions of Prior Beliefs



Left panel: model implied weights on current and future signal as functions of prior mean beliefs on the future state. Right panel: model implied standard error of the noises attached to current and future signal. The blue curves are for future (SPF) signal and red curves are for current signal.

Figure 17 shows that the optimal choices of weights (left panel) and precision (in terms of standard error, right panel) indeed depend on the agent's prior belief. In particular, when the prior belief leads to on average a bad state in the future, the agent will first pay more attention to the current signal, then shift to SPF signal as the implied state getting worse. This piece of evidence is also consistent with my empirical finding. As the prior mean is accumulated from the history of signals and usually not observable in the data, it can then be thought of as a proxy of the "internal state" in my empirical section. As discussed in Section 4.2.4, both the current state of the economy and the internal state accumulated from the past signals play a role in creating the state-dependent marginal effects of signals.