Learning and Subjective Expectation Formation: A Recurrent Neural Network Approach

Chenyu Hou *

August, 2020

Version 2.2

Abstract

Households are exposed to a rich set of signals that can help them form expectations. How they make use of these signals remain a debated question among economists. Standard learning models rely on structural assumptions and parametric methods that are typically incapable to capture potential non-linear and state-dependent relationship between signals and expectations. This paper tackles the problem by proposing a generic learning model that can cover a large class of expectation formation models, including those are standard in the literature. The average structural function of this model is estimated with an innovative semi-parametric approach: Recurrent Neural Network. Average marginal effect of signals on expectational variable is estimated using Double De-biased Machine Learning estimator, together with valid inferences. Applying this approach to survey expectations for U.S. households, I find: (1) agents' perceptions about future economic condition have asymmetric and non-linear response to signals; (2)the attentions to past and future signals in the learning model are highly state-dependent, agents are adaptive learner in ordinary periods and become forward looking as state of economy gets worse; (3) both signal and exposure to news on economic condition play important role in creating the attention-shift. I then propose a model that features rational inattention to explain these patterns.

*Hou: Vancouver School of Economics, University of British Columbia.

1 Introduction

Models on expectation formation play an important role in modern macroeconomic theories. For example, the workhorse DSGE model is built on full information rational expectation (FIRE) assumption following Lucas (1976). However agents' optimal choices under FIRE may be very different from those with imperfect information or other forms of bounded rationality [1]. These frictions may also dampen or amplify the effect of policies and economic shocks. For these reasons various types of expectation formation models have been examined by researchers in the past decade. These models typically assume households form beliefs about future upon observing signals on state of economy. One of the key questions economists have been asking is then how these signals affect households' expectation.

To answer this question, empirical studies typically try to look at survey data on household expectations through the lens of these models. Parametric tools are then applied to the data using specific functional forms implied by a structural model on expectation formation. Findings with this approach are subject to parametric assumptions made by researchers thus making the empirical findings model-specific. For example, in a standard noisy information model [2] we will find signal with the same magnitude has the same impact on expectation as long as agents have the same prior mean. However with models where agents learn about structural parameters recursively[3] or features regime switching, we will find the impact of the same signal changes at different point of time.

In this paper, I also explore how signals affect households' expectations but without imposing parametric assumptions and specific functional forms. Instead I try to recover the functional form of expectation formation model from data directly using a semi-parametric method, Recurrent Neural Network (thereafter RNN). I propose a generic learning model that is flexible about structure that households employed to form expectation, while maintaining some minimal assumptions. Specifically, the model considers that households form

[1] For example, see Caplin (2016) for a detailed discussion.

[2] In these models agent forms expectation on a single unobservable state using stationary Kalman Filter. See examples as Coibion and Gorodnichenko (2015),Coibion and Gorodnichenko (2012) and Andrade and Le Bihan (2013). For noisy information model allowing for joint expectation formation, see *Hou (2019)* and Kamdar (2019).

[3] For instance Malmendier and Nagel (2015).

expectation about future by perceiving some latent variables according to a rich set of signals they observe. Such an assumption is common in a large class of structural learning models, but it may take different forms depending on assumptions in the model. For example, in standard noisy information model the latent variable agents perceive is the posterior mean of state that is not perfectly observable by them; and in Hidden Markov Models these latent variables become their posterior beliefs on probability of the Markovian state.

The difference between this paper with the others is that I imposes no restrictions on what these latent variables are, how the signals affect these latent variables, and how the latent variables affect households' expectations. The mapping from signals to expectational variables through the latent structure will be recovered by RNN from the data observed. The strength of this approach is that it can capture the correct relationship between signal and expectational variables that are generated by any learning model described in the class above. In particular, if macroeconomic signals affect expectation non-linearly, or in a fashion that changes across time, these relations will be captured by RNN but 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[4].

After approximating the unknown functional form of expectation formation model with RNN, I estimate the average marginal effect of macroeconomic signals using the approximated model. To assess the statistical significance of my findings, double de-biased estimators on marginal effects are obtained following the approach by Chernozhukov *et al.* (2018).

Applying this method to Michigan Survey of Consumers, I document three major findings. I first show households' expectations on unemployment and economic condition are non-linear functions of signals about the corresponding subject - the effect of an incremental change in such a signal depends on the level of itself. The relationship is also asymmetric - positive and negative signals with same magnitude have asymmetric impact on expectations, households respond more aggressively to signals that suggest bad state of economy.

Using the approximated functional form of expectation formation model, I then find the

[4] In **Appendix B.4** I illustrate these properties with examples using simulated data.

marginal effects of these signals change along time. Specifically, marginal effects of signals about realized economic conditions fall as GDP growth slows down or unemployment rate hikes up, however the opposite is true for signals that contain information about future. When interpreting marginal effects as weights that households put on signals, this finding suggests households shift their attentions from signal about current and past states to those about future. In other words, households are "adaptive learners" when economic conditions are stable and become more "forward looking" when situation gets worse.

Lastly the approximated generic learning model suggests such attention shift is mainly driven by signals on economic conditions and exposure to news on these topics, rather than information related to interest rate or inflation. And it contributes through both the contemporaneous signals newly observed in each period and the latent states that capture the impacts of the signals happened in the past.

This paper contributes to three strands of literature. It first relates to the growing empirical literature using survey data to investigate how expectations are formed. These studies have documented substantial amount of evidence on agent's expectation deviating from Full Information Rational Expectation (thereafter FIRE) benchmark. For example, Coibion and Gorodnichenko (2012) and Andrade and Le Bihan (2013) show that expectations are formed under a limited information structure that is at odds with FIRE assumption. However this literature is silent about possibility that agents' expectation formation process is non-linear and state-dependent.My paper helps to fill in this gap. My findings first verify that information is rigid in household's expectation as signals they observed in the past have substantial impact on their expectations today. I then add new evidence to the literature that household's expectation formation model is non-linear and asymmetric, and the informational weights they put on signals depend on the state of economic conditions. These findings are simply not available in empirical analysis motivated by class of linear models.

This paper is also related to literature on learning and information transmission. In this paper I examine how households form expectations using information available to them. What information households use to form expectation is particularly important in policy making and explaining business cycle fluctuations. Households can acquire information from realized signals such as current and past inflation or unemployment status. This type of agents are adaptive learners, an example is Evans and Honkapohja (2001) where households use only realized fundamentals and a least square approach to form expectations.Households can also get signals that contain information about future, for example, from news. Carroll (2003) first shows evidence that information about future transmit from Professional Forecasts to households' expectation through news media. And more recently Barsky and Sims (2012) shows that households' expectation on economic conditions from Michigan Survey of Consumers contains information about future. Following Carroll (2003), I include both Professional Forecasts and realized macroeconomic fundamentals as signals and I find whether households are adaptive learner or acquire more information from professionals, thus being forward looking, depends on the status of economy. They rely more on signals about fundamentals when economic conditions are stable and pay more attention to views from professionals when economy deteriorates and there are more news coverage. Such a finding is consistent with the view from rational inattention models: agents may allocate their limited resources to different information sources depending on the status of economy they perceived.

The methodology part of this paper contributes to the recent literature using machine learning techniques to solve economic 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)[5]. Among these tools the use of deep Neural Networks in macroeconomics can date back to early 2000s. Back then different types of Neural Networks (both fully-connected ones and recurrent ones as used in this paper) were used to forecast economic variables such as inflation and exchange rate. Examples like Nakamura (2005) and Kuan and Liu (1995) have shown RNN out-performs standard linear models in forecasting inflation and exchange rate respectively[6]. However, this paper uses RNN to solve both prediction and estimation problems. RNN is first used to approximate the average structural function (henceforce ASF) as described in Blundell and Powell (2003) derived from generic learning model, which is essentially a prediction problem.

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

[6]See also Almosova and Andresen (2018), R. and Hall (2017) for recent application to macroeconomic forecasting with state-of-art architecture of RNN – Long Short Term Memory (henceforce LSTM) layers.

Then with the standard identification restrictions same as empirical literature on learning and expectation formation, I follow Chernozhukov *et al.* (2018) and use their semi-parametric moment condition model to get double-debiased estimator and its inference for the average marginal effect of signals on households' expectations. The estimation procedure of this paper is closely related to those in Chernozhukov *et al.* (2018) and Farrell *et al.* (2018). The latter offers convergence-speed conditions for deep Neural Networks to acquire valid inference. To my best of knowledge, this is the first time RNN is applied to learning and expectation formation problems in an estimation context.

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

2 Generic Learning Model

In this section, I describe a conceptual model about how expectation is formed by households, which I refer as Generic Learning Model. It is worth to describe the similarity and key differences between this model to 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 behaviour. These parametric assumptions include both the underlying structure agents learn about and the way learning is carried out. For example, in standard noisy information models the perceived law of motion agents learn is assumed to be linear, and the prior and posterior process are structured as Gaussian. These assumptions lead to specific parametric regression methods used in different learning models. The Generic Learning Model maintains standard assumptions on information structure and identification but impose only minimal restrictions on functional forms of learning. This then naturally requires the use of non-parametric or semi-parametric methods such as RNN. Such a feature also implies the Generic Learning Model has the ability to represent a large class of learning models existing in the literature despite these models may differ in their functional forms. In **Appendix B.4**, I include an example that illustrates how the Generic Learning Model formulation can represent a stationary Kalman Filter.

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

2.1 Agent's Problem

Consider agents observe 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 public signal will be official statistics such as CPI inflation or professional forecast on CPI inflation a year from now. An example of private signal will be state-level inflation matched to the location agent lives at or the fraction of news stories about inflation published on local newspapers.

Other than public and private signals, there is an individual level noise term denoted as $\epsilon_{i,t}$ in 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). More generally it can also stands for any unobserved individual level information that is not captured by public and private signals but is used by the agent when forming expectation. If it takes form of observational noise, $\epsilon_{i,t}$ is typically separable additive to the public and private signals. Here for generality I do not restrict the form it enters expectation formation process.

After observing the set of signals $\{X_t, s_{i,t}, \epsilon_{i,t}\}$, agent forms expectation of variables Y_{t+1}

and denote the corresponding subjective expectation as $Y_{i,t+1|t}$ [7].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 the most 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 generic learning model:

Assumption 1. *Agent forms expectation through two steps: updating and forecasting. In updating step, agent forms 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 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 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 expectation to be affected by signals in the past without explicitly specifying a fixed length of memory[8].

[7] 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.

[8] For example, one may want 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)

These two steps are commonly seen in standard learning models. For example, in stationary Kalman Filter, this is the usually referred as "Filtering Step" where agent uses the new signals to form a "Now-cast" variable about the current state of economy. They will then use this "Now-cast" to form expectation about future using their perceived law of motion. [9] This step is the same as "forecasting step" in generic learning model.

It is then worth noting that the structure of generic learning model described in assumption 1 covers a large class of learning models existing in the literature, other than the stationary Kalman Filter. It is obvious this formulation includes adaptive learning models where agent uses only past information to form expectations[10]. It also covers models where agents get information about future from professional forecasts through reading news stories, as in Carroll (2003). Recall the $\hat{\mathbb{E}}$ in equation (1) means agents may form expectation using subjective beliefs, instead of assuming the full structure of agents' knowledge and the statistical property agent believes in as usually done in the learning literature. This allows for behavioural models such as Bordalo *et al.* (2018). To further illustrate the generality of this model, in Appendix B.4 I will take the stationary Kalman Filter that is typically used in noisy information models as an example and represent it in the form of the generic learning model.

In addition to Assumption 1, I also need independence assumptions on the observational noise term $\epsilon_{i,t}$. This assumption states the noise unobservable by economists is independent with observed public and individual specific signals. And it's independent across individuals and time. While such an assumption is commonly made in noisy information as well as other learning models with unobserved noise, the economic intuition behind it is simple as well. Consider 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 price change only for one or several products. Mathematically this signal can be thought of as drawn from a distribution, with official measure of inflation being the mean of this distribution. An individual may accidentally draw the signal from left tail or right tail of the distribution, depending on the specific product he or she picked up. The public signal X_t

[9] Refer to Appendix B.4 for a detailed example in the context of standard noisy information model.
[10] See Evans and Honkapohja (2001) for examples.

(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 from its mean as well as across individual and time.

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

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

The flexible form of expectation formation in (1) together with the two assumptions summarize the generic learning model.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 $\Theta_{i,0}$ is observable[11].

2.2 Economist'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 household's expectations. In standard learning literature, this is achieved by making structural assumptions on the expectation formation process, e.g. the functional forms of $F(.)$ and $H(.)$ and estimate the average marginal effect of signals or structural parameters through parametric methods. The findings from this approach are model-specific and prone to model mis-specification. An alternative way to estimate average marginal effect is through estimating the Average Structural Function(ASF) without imposing assumptions on the form of $F(.)$ and $H(.)$. And use the ASF as a nuisance parameter to estimate average marginal effect.

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

[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 $F(.)$.

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 $\epsilon_{i,t}$:

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

In (4), function $\mathcal{F}_{\epsilon}(.)$ is the marginal 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 of 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 contrary if the estimated ASF is highly non-linear it suggests there is 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 $G(.)$, with information we have. We then want to assess how signals affect households' expectations. In nonparametric models 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 properties of these structural functions. In particular, the "average derivative" of ASF can be an important measure for marginal effects of input variables, in this paper I define such a derivative as average marginal effect of signals on expectations. The goal now is to estimate the ASF as well as the average marginal effect of the Generic Learning Model.

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 and the dimensionality is unknown.

In standard learning literature, these problems can be solved by parametric assumptions on structural function. For example in models with explicit form on forecasting and updating steps, such as stationary Kalman Filters, $F(.)$ and $H(.)$ are parametric functions and parametric regressions can be applied to the reduced form relation between expectational variables and signals. This method will lead to a "best estimate" of ASF within the set of models that satisfy the parametric assumptions. In this paper I take an alternative approach where I directly estimate the ASF with semi-parametric methods – Recurrent Neural Network. Later using the estimated ASF as first-stage nuisance parameter, I construct a second-stage double debiased estimator of the average marginal effect following Chernozhukov *et al.* (2018). I start by introducing the RNN approach to directly estimate the Average Structural Equation.

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 by their ability to approximate any functional forms between input and output variables. Such a property is implied by the Universal Approximation Theorem addressed in Hornik *et al.* (1989), which suggests a single layer neural network with sigmoid activation function can approximate any continuous function. However, the most popular Feedforward Neural Networks don't fit the problem well because of its lack of ability to model time dependency between output variables and past input variable. To solve this problem, the Recurrent Neural Networks are used.

RNN are neural networks designed to model time-dependency between input and output variables. When the mapping between input and output variables are described by a dynamic system, it is shown by Schäfer and Zimmermann (2006) that RNN can approximate dynamic systems of any functional form arbitrarily well. This is usually referred as the Universal Approximation Theorem for RNN. It turns out that the ASF of Generic Learning Model takes the form of dynamic system considered by this Universal Approximation Theorem.

This justifies why the ASF (4) can be estimated by Recurrent Neural Networks.

For the ASF to be represented in form of dynamic systems, I need the assumptions 1 and 2 that $\Theta_{i,t}$ is a Markov Process and $\epsilon_{i,t}$ is i.i.d across individual and time.Theorem 1 shows that the ASF (4) 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. There exists finite dimensional $\theta_{i,t} \in \mathbb{R}^d$, continuous function $f : \mathbb{R}^d \to \mathbb{R}^l$ and 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 well-approximated by a dynamic system:*

$$\begin{aligned} y_{i,t+1|t} &= f(\theta_{i,t}) \\ \theta_{i,t} &= h(\theta_{i,t-1}, Z_{i,t}) \end{aligned} \tag{5}$$

The Average Structural Function can then be denoted as:

$$y_{i,t+1|t} \equiv g(\{Z_{i,\tau}\}_{\tau=0}^{t}, \theta_{i,-1}) \tag{6}$$

Notice equation (6) is an alternative representation of ASF (4). In (6) 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 $t = 0$, $\theta_{i,-1}$. 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 measure for wellness of approximation and proof of Theorem 1 can be found in **Appendix A**.

Following Schäfer and Zimmermann (2006), Recurrent Neural Network (RNN) is the universal approximator of dynamic system in forms of (5).[12] Theorem 1 then implies I can use a state-of-art RNN with Rectifier Linear (ReLu) activation function to approximate the ASF (4) derived from the Generic Learning Model.[13] Within the class of functions in RNN $\mathcal{G}_{f\circ h}^{RNN}$, the estimator is computed by minimizing the sample mean squared errors:

[12]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).

[13]The RNN approximate dynamic systems (5) by constructing representations of $\theta_{i,t}$ as well as $f(.)$ and $h(.)$. For introduction to a simple RNN and how the class of functions in RNN is determined by choices of its architecture, refer to **Appendix X**.

$$\hat{g}_{rnn} := \underset{g_w \in \mathcal{G}_{foh}^{RNN}}{\arg\min} \sum_{i,t} \frac{1}{2} \left(y_{i,t+1|t} - g_w(\{Z_{i,\tau}\}_{\tau=0}^{t}) \right)^2$$

In Theorem 1 the alternative representation (6) 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 allow for endogenous time-varying marginal effect of signals $Z_{i,t}$. This point will become more clear when I introduce 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 $g(\{Z_{i,\tau}\}_{\tau=0}^{t})$:

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

Or for a single signal z^j which is the $j-th$ element in Z, this can be written as:

$$\beta^j = \mathbb{E}[\frac{\partial g(Z, \theta_{-1})}{\partial z^j}] \tag{8}$$

The equation (7) can be thought of as a moment condition that is 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) contains a biased term which is not centred around zero and diverges in general; (2) the estimate is usually not well-behaved asymptotically which makes inference hard.[14]

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

One way to solve these problems is to use the Double De-biased Machine Learning (DML) Estimator 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, β; infinite dimensional nuisance parameter η 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 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.[15].

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

$$\mathbb{E}[\beta^j - \frac{\partial g(Z,\theta_{-1})}{\partial z^j} + \frac{\partial ln(f_z(Z,\theta_{-1}))}{\partial z^j}(y - g(Z,\theta_{-1}))] = 0 \tag{9}$$

The nuisance parameters associated with moment condition (9) then include both the average structural function $g(.)$ as well as the joint density function $f_z(Z,\theta_{-1})$. 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 to estimate the joint density of $f_z(Z_{i,t}, Z_{i,t-1})$. The joint density is then obtained using multivariate Gaussian Kernel Density Estimation with bandwidth chosen according to Silverman (1986).

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.

2. Obtain estimate of conditional expectation function from computing derivative numer-

[15] 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**

ically:

$$\frac{\partial \hat{g}_{rnn}}{\partial z^j_{i,t}} = \lim_{\delta\to 0} \frac{\hat{g}_{rnn}(X_{i,t}+\Delta_j/2, \{X_{i,\tau}\}^{t-1}_{\tau=0}, \theta_{-1}) - \hat{g}_{rnn}(X_{i,t}-\Delta_j/2, \{X_{i,\tau}\}^{t-1}_{\tau=0}, \theta_{-1})}{\delta}$$

Where $\Delta_j \in \mathbb{R}^{d_1+d_2}$ is a vector of zeros, with jth element being δ.

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.

$$\frac{\partial ln(\hat{f}_z(Z,\theta_{-1}))}{\partial z^j_{i,t}} = \lim_{\delta\to 0} \frac{\hat{f}_z(Z_{i,t}+\Delta_j/2, Z_{i,t-1}) - \hat{f}_z(Z_{i,t}-\Delta_j/2, Z_{i,t-1})}{\delta \hat{f}_z(Z_{i,t}, Z_{i,t-1})}$$

4. Then the DML estimate is given by:

$$\hat{\beta}^j = \frac{1}{N}\sum_n \frac{1}{T}\sum_t \underbrace{[\frac{\partial \hat{g}_{rnn}(\{Z_{i,\tau}\}^t_{\tau=0}, \theta_{-1})}{\partial z^j_{i,t}} - \frac{\partial ln(\hat{f}_z(Z,\theta_{-1}))}{\partial z^j_{i,t}}(y_{i,t} - \hat{g}_{rnn}(\{Z_{i,\tau}\}^t_{\tau=0}, \theta_{-1}))]}_{\equiv \hat{\beta}^j_{i,t}}$$

4 Application to Survey Data

In this section I use survey data of expectation and a rich set of macroeconomic signals to estimate the generic learning model. There is a growing literature using survey data to estimate learning models. For example, Coibion and Gorodnichenko (2015) and Andrade and Le Bihan (2013) use survey data in framework of noisy information model and information rigidity models; and Malmendier and Nagel (2015) use survey data to estimate a Least Square Learning model with a time-varying decay of macroeconomic signals observed.

The respondents in the survey researchers use are different as well. The most widely explored expectations are those from households and professionals. In this paper I focus on household's expectation from US and I use professional forecasts as a signal that household can utilize to form their expectations.

4.1 Data Description

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

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: $E\pi_{t\|t-1}$ change of economic condition: $E\Delta y_{t\|t-1}$ unemployment change: $E\Delta u_{t\|t-1}$ interest rate change: $E\Delta r_{t\|t-1}$	Michigan Survey of Consumers
Output variable ($\hat{Y}_{i,t+1\|t}$)	**Variable and Notation**	**Source**
Expectational Variable	inflation rate: $E\pi_{t+1\|t}$ change of economic condition: $E\Delta y_{t+1\|t}$ unemployment change: $E\Delta u_{t+1\|t}$ interest rate change: $E\Delta r_{t+1\|t}$	Michigan Survey of Consumers

For outcome variable $\hat{Y}_{i,t+1|t}$ 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 respondent's one-year-ahead expectation on various macroeconomic aspects. In this paper I include four expectational variables of interest: (1) expected inflation rate, denoted as $E\pi_{t+1|t}$; (2) whether economic condition will be better, denoted as $E\Delta y_{t+1|t}$; (3) whether unemployment rate will increase, denoted as $E\Delta u_{t+1|t}$; (4) whether interest rate will increase $E\Delta r_{t+1|t}$.[16]

I include two sets of public signals X_t. One is the realized economic statistics from Federal Reserve of St. Louis. These signals contain information about the current state of the economy. In the adaptive learning literature agents rely only on (the history of) this information to form forecasts. Another set of public signals I consider are the professional forecasts from 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.

Then in individual level signals $S_{i,t}$, I include local unemployment rate and CPI matched with individual in MSC according to their location information. I also include intensity of news story reports on recessions, inflation and interest rates at both local and national level[17]. 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[18]. 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 have two waves for each individual, whereas

[16]The exact questions and interpretations corresponding to these variables can be found in **Appendix XXXX**.

[17]I scraped volume of reports on corresponding macroeconomic topics from TV news scripts and local newspaper articles. Following PFAJFAR and SANTORO (2013) I construct the news exposure measure as total volume of news on each topic (for example, news about inflation) in each quarter as a fraction to total news 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.

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

capturing the latent state accumulated by observing history of signals requires 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, which are the four characteristics found significantly affecting expectation by **?**. The baseline sample I'm using is quarterly from 1988 quarter 1 to 2019 quarter 1. The length of sample is due to the availability of data on news stories[19]. 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.* (2018) 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 $K = 6$.[20] Table 2 summarizes the architecture of RNN I use.

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

[20] 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 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 impact of signals on expectations regarding the same subjects, which I refer to as "self-response". For example, impact of realized unemployment rate on expectation about unemployment for the future. Another interesting direction is to examine "cross-response", for example, how signals on inflation affect unemployment expectation. This will help us to understand how households believe different economic aspects interact with each other. Results on these topics are documented in details in my other work *Hou (2020)* thus are not included in this paper.

Because 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 Average Structural Function from the baseline RNN described in Table 2. Then I present the time-varying marginal effects of macroeconomic signals implied by the estimated ASF to illustrate the key finding that households are adaptive learners in ordinary periods and become more "forward looking" when economic conditions get worse. I interpret this finding as an "attention shift" of households from signals about past and current state of economy to signals that contain information about future. After that I obtain 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.

4.2.1 Estimated Average Structural Function

First it's worth examining the estimated Average Structural Function. As the ASF in (4) is a complex object with high-dimensional input and 4-dimension output, it's hard to visualize such a function in all possible dimensions. I decide to focus on presenting expectation as function along one dimension of the input signal as a starting point because it serves as a foundation to understand the results presented in this section. Before I plot the function in a two-dimension space, it's useful to define the estimated function in that space. 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 (5), the estimated functional estimator can be expressed as the following function

$$E_t x_{i,t+1} = \hat{g}_x(\theta_{i,t-1}, Z_{i,t}^{-x}, x_t) \tag{10}$$

Now take unemployment as an example subject. Figure 1 plots the average structural function of expected probability for unemployment rate increase, along the signal on change of actual unemployment rate. Following (10), this function can be written as:

$$E_t \Delta u_{t+1} = \hat{g}_u(\theta_{i,t-1}, Z_{i,t}^{-u}, \Delta u_t) \tag{11}$$

In Figure 1, the function (11) is plotted at three different points of time: quarter 2,3 and 4 in 2016, and different from each other. However such a difference is not due to estimated functional form $\hat{g}_u(.)$ is different across time, but because of different inputs of $\theta_{i,t-1}$ and $Z_{i,t}$[21]. This means at different point of time households may form different expectations in

[21]Given that they are close to each other in time (should have similar hidden state accumulated) and current Δu_t is roughly at 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 falling gradually illustrate how expectation is slowly adjusting downwards when actual unemployment rate keep falling($\Delta u_t <$ 0) throughout the three quarters plotted.

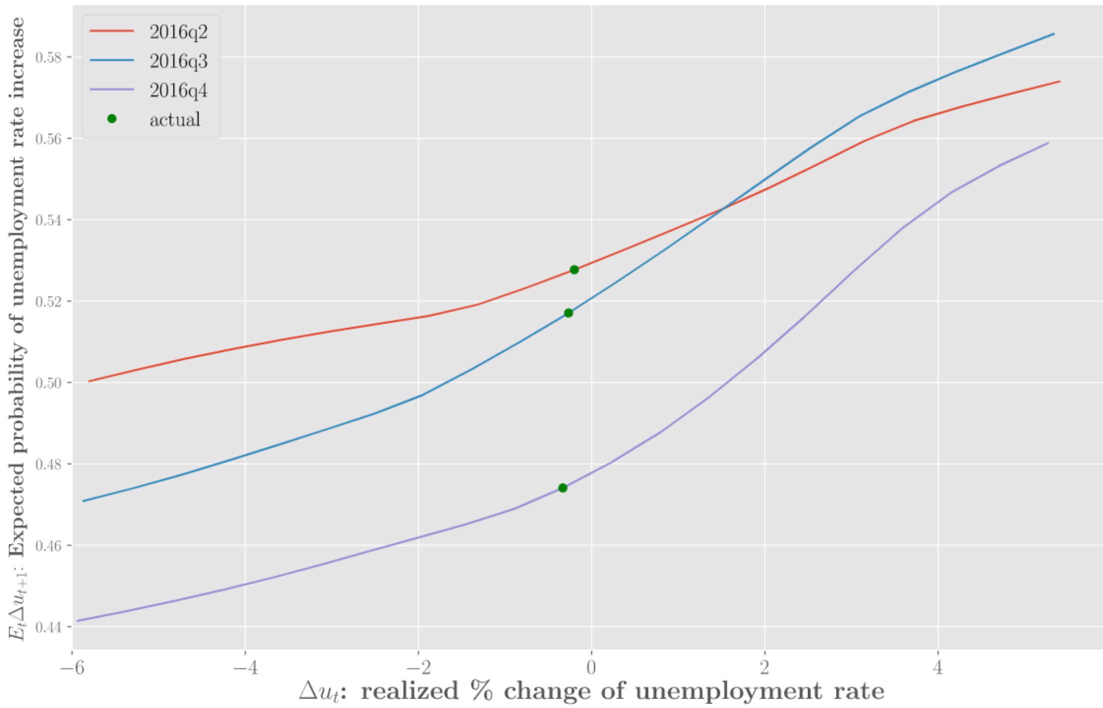


Figure 1: Average of expected probability for unemployment rate increase $E_t\Delta u_{t+1}$ as function of realized unemployment rate change Δu_t, at different point of time. Purple curve: 2016q4, blue curve: 2016q3, red curve: 2016q2. The dot on each curve represents the prediction from estimated function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

response to the same signal on realized unemployment rate change, but such a difference comes from either the hidden state ($\theta_{i,t-1}$) they accumulated from observing a different path of signals, or the interactions between newly observed signals $Z_{i,t}$. In other words, any state-dependency I find with the estimated ASF is endogenously resulted from 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.

Then I turn to the properties of estimated ASF. A first thing to notice is that it's highly non-linear and it is common to all three points of time despite the potential difference between hidden states and covariates. In particular the slope of ASF changes in three regions. First when actual changes of unemployment rate is negative and big in absolute value, for example less than 0% or lower in Figure 1, the slope of ASF is relatively flat. This means households are less sensitive to those information, which can be thought of as "good" news. The the slope gets steeper and steeper in the region $0\% < \Delta u_t < 2\%$, which means agents' expectation become more responsive to the actual unemployment status when unemployment rate increases. Finally when unemployment rate increased too much that Δu_t becomes higher than 3% or more, the slope of ASF becomes flat again.

A second important observation of Figure 1 is that the ASF is asymmetric. Taking 2016 quarter 4 as example, which is the purple curve in Figure 1. In that quarter unemployment rate decreased by around 0.4% and on average people expect unemployment rate to increase a year from now by probability 0.45. Keeping other signals (and the history of them) fixed. The curve implies if unemployment had increased by 1.6% instead of falling by 0.4%, the model predicts households will be 5% more likely to believe unemployment will increase in the future. However, if unemployment decreased further by 2.4%, households will only be 3% less likely to expect unemployment rate going up. This implies households may be more sensitive to news that is unfavourable, unemployment rate increase in this case. Such a pattern won't be seen in a linear model[22]if the underlying expectation formation model is linear in signals, the ASF will be linear as well.

A final point to notice is because of the time-variation of latent state $\theta_{i,t-1}$ and covariates $Z_{i,t}$, the slope of ASF becomes time-dependent. This gives rise to the time-varying marginal effects of signals. I will discuss the details on this in **Section 4.2.2**.

Now I have showed you the estimated ASF is non-linear and asymmetric with fixed input $\{Z_{i,\tau}\}_{\tau=0}^{t}$, as it is still an estimated object it's useful to get a sense of how significant these patterns are. To achieve that I turn to estimate average deviations of expectation and obtain valid inference using DML as described in **Appendix B.1**

$$\gamma_\delta = \mathbb{E}[g(Z_{i,t}+\delta, \{Z_{i,\tau}\}_{\tau=0}^{t-1}, \theta_{i,-1}) - g(Z_{i,t}, \{Z_{i,\tau}\}_{\tau=0}^{t-1}, \theta_{i,-1})] \tag{12}$$

The average deviation is defined in equation (12), 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 focus on the pairs in which output expectational variable and input signal variable are on the same subject.

In Figure 2 I plot again average deviation for unemployment expectation along the change of unemployment signal as a leading example. I consider 20 different values of δ symmetrically centred around 0. And for each point estimate at δ, I present the 95% confidence interval. It

[22]See discussion in Appendix B.4

shows the average deviation exhibits same curvature as the shape of estimated ASF presented in Figure 1. It indicates the responsiveness of expectation on unemployment status to realized unemployment signal is relatively weak when unemployment rate falls or increases by large magnitude, meanwhile the expectation is most sensitive to unemployment signal when the signal increases but by smaller magnitudes. The confidence interval shows the asymmetry is significant. With a positive change to Δu_t, expected unemployment will increase more in absolute value comparing to how much it decreases in response to a negative change of Δu_t with same magnitude, and such a difference is statistically significant.

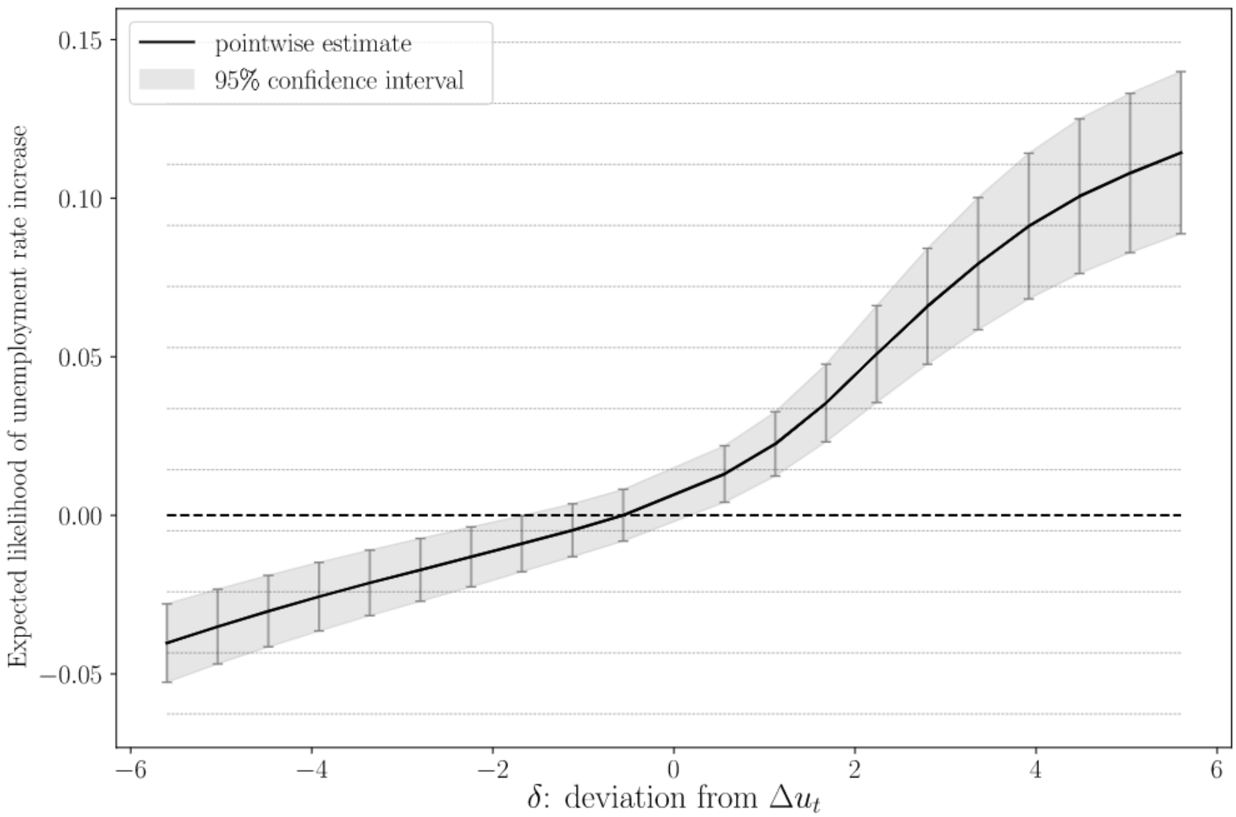


Figure 2: Average change of unemployment expectation $E_t\Delta u_{t+1}$, when unemployment signal Δu_t change by δ. This is obtained by point-wise estimation of (12) at 20 different points of δ, following the Double-Debiased Machine Learning approach from Chernozhukov *et al.* (2018), 95% confidence band is reported at each point-wise estimate. The estimate is depicted as solid black line, shaded area is implied 95% region.

If I look at the average deviation for all four expectational variables, and again focus on the "self-response". I find such a non-linearity shows up consistently in cases of unemployment expectation and economic condition expectation. This can be seen from comparison between panel (a) and (b) in Figure 3. In panel (b) when Δy falls drastically the slope of ASF becomes flat, same as the case when unemployment signal is high in panel (a). Then it gets steeper as δ becomes closer to 0, and gets flatter again when δ keeps increasing and becomes positive. On the other hand, in panel (c) and (d) of Figure 3 which correspond to inflation and interest rate expectation as function 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 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'll focus on presenting results with expected economic condition, $E_t\Delta y_{i,t+1}$, from now on. For results on other three expectational variables I include the results in **Appendix C.1**.

4.2.2 Time-varying Marginal Effect

Following from the estimated ASF in (10), I can define the average (across individual) time-specific marginal effect of signal x on expectational variable Ex as:

$$\beta^{Ex}_{x,t} = \mathbb{E}_n[\frac{\partial \hat{g}_x(\theta_{i,t-1}, Z^{-x}_{i,t}, x_t)}{\partial x_t}] \tag{13}$$

This marginal effect is different at each point of time t for the same reason as discussed in **Section 4.2.1**: 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 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.[23] In this section I show profound time-

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

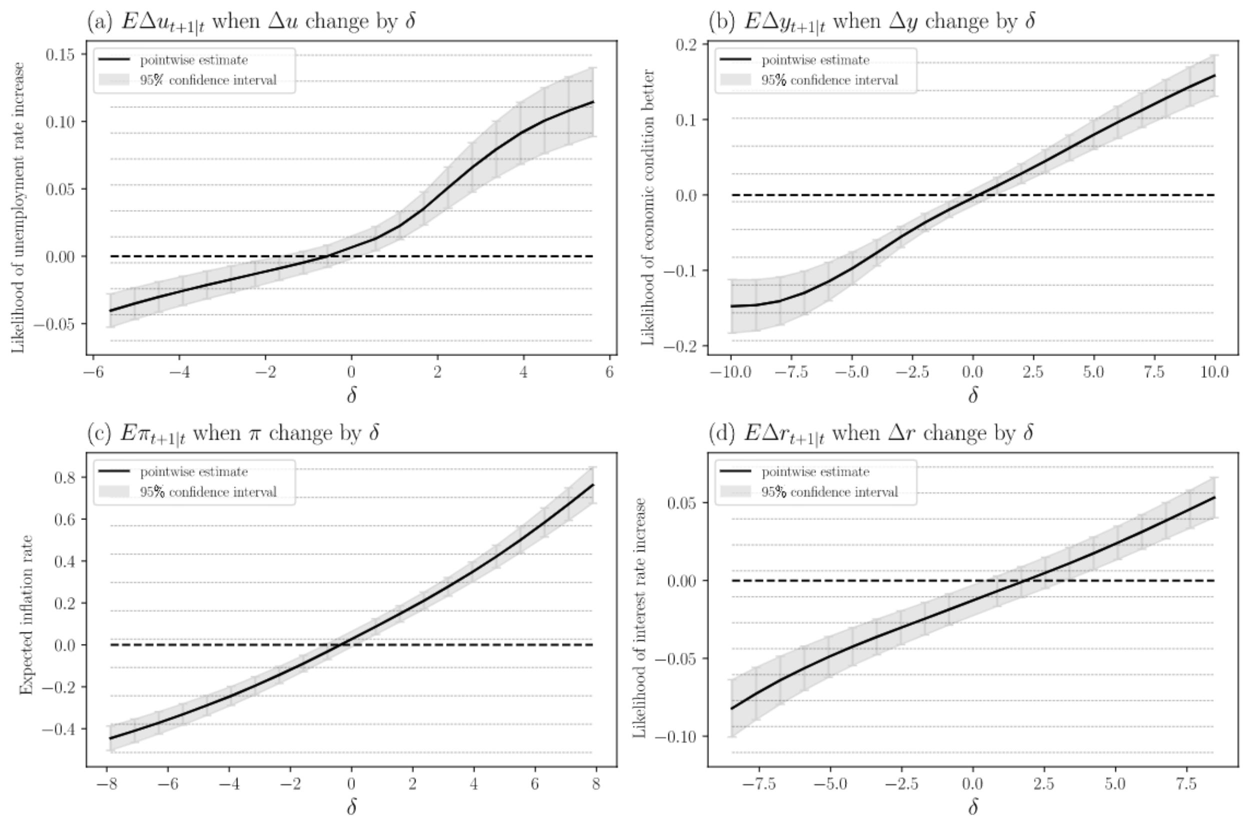


Figure 3: Average deviation of four expectational variables in response to signals on themselves. Panel (a): expected likelihood of unemployment increase as unemployment signal change by δ. Panel (b): expected likelihood of economic condition be better as real GDP signal change by δ. Panel (c): expected inflation rate as inflation signal change by δ. Panel (d): expected likelihood of interest rate increase as interest rate signal change by δ.

variation in the average marginal effect of signals on expectations about economic condition. Specifically, such a time-variation implies households' attention to signals are cyclical: they put less weights to signals about current and past states and at the same time more weights on signals about future, during periods with bad economic conditions. In other words, when agents form expectation about economic conditions, they change from adaptive learners to forward-looking during bad time.

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

Signals about past v.s. future: Following the adaptive learning literature, agents can acquire information about current state of economy from macroeconomic statistics. They get

different in levels at each point of time.

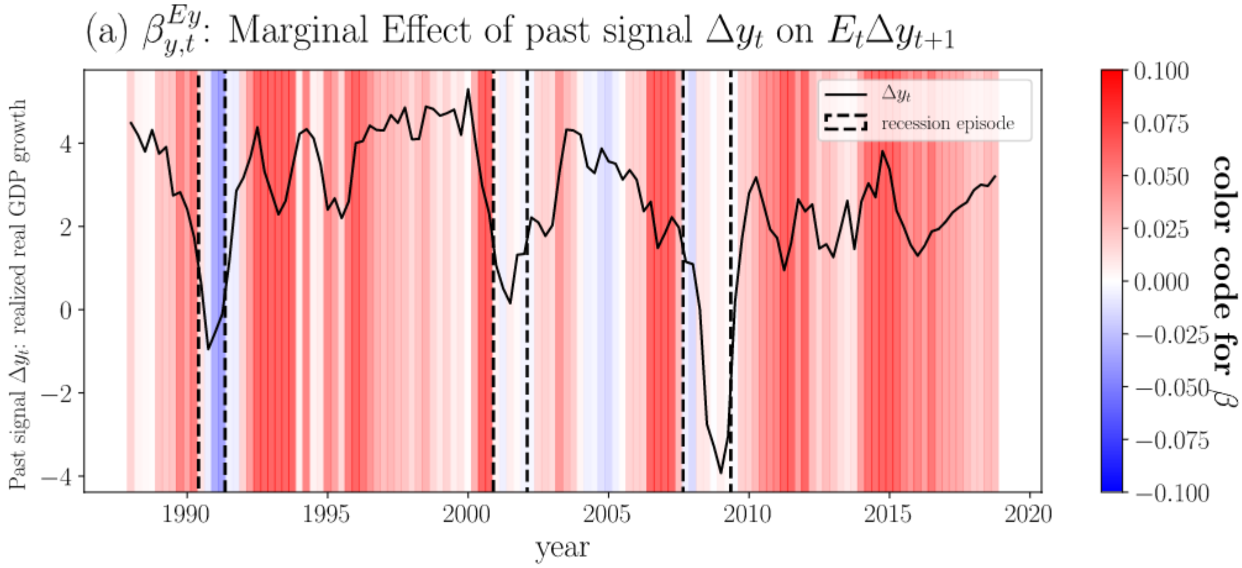
this information either directly as it's public available, or partially through daily activities. I will use realized key macroeconomic variables as proxy for signal about past, expectation formed majorly relying on this information is then treated as adaptive. For signals about future, I follow Carroll (2003) and use consensus (average) expectation from Survey of Professional Forecasters as proxy. Information about future can take the form as news or anticipated shocks as in Beaudry and Portier (2006) and Barsky and Sims (2012) and it flows into 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 unemployment rate increasing[24]:1990q3-1992q3,2001q1-2002q4 and 2007q3-2010q3. This is because I use year-to-year change of unemployment rate as 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 is 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 end date of NBER recessions when they form expectations around the same time[25]. The results won't change qualitatively if I use the NBER recession dates as measure of "bad time", these results are included in **Appendix C.2**.

I then present the time-specific marginal effect from (13) of signals on real GDP growth. I consider both signals about past and future. In Figure 4, the color bars in top panel are the marginal effects of real GDP growth signal, $x_t = \Delta y_t$, on expected economic condition next year; those in 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.

[24]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 $\Delta u_t > 0$

[25]The announcement typically comes out at least 2 quarters after the official end day of NBER recession.



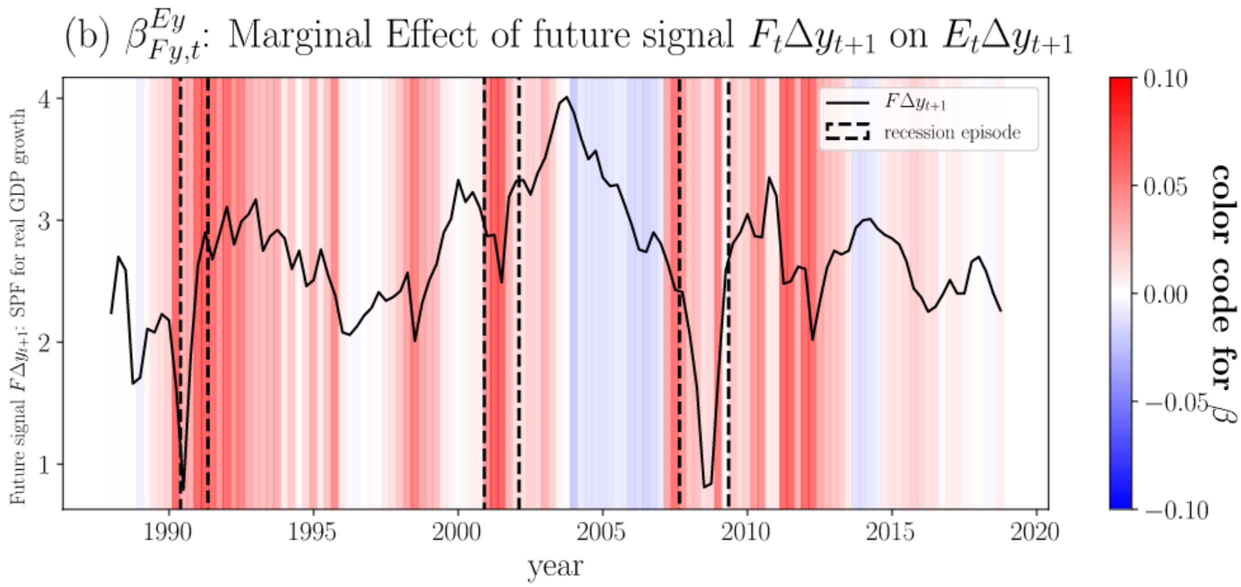


Figure 4: Color bars in panel (a): the marginal effects of real GDP growth signal Δy_t on expected economic condition next year $E\Delta y_{t+1|t}$. Panel (b): the marginal effects of professionals' forecasts about real GDP growth next year $F\Delta y_{t+1|t}$ on expected economic condition. Red color: positive marginal effect; blue color: negative marginal effect. Black solid line: data on frequency of news about recession.

The color bars in each panel stand for the corresponding marginal effect at that point of time. A red color means positive marginal effect, the blue color means negative marginal effect and white means the marginal effect is zero. The color map is on right side of each panel, the scale stands for normalized marginal effect. For example, 0.1 on the color map means when signal x_t change by 1 standard deviation, corresponding expectation changes by 0.1 standard deviation. This is then represented by a dark red color bar in the graph.

The darker the color, the bigger the magnitude for marginal effect. The solid black line is the series of signal x_t at which I evaluate marginal effect. And the dotted area is NBER recession episode.

In general, both higher real GDP growth and higher forecasted growth by professional 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 0.24 standard deviation increase in expected business condition (on average 0.125 more likely to believe economic condition to be better).

One key observation comes from comparing top panel to bottom. In panel (a) the pale color during recession periods in panel (a) suggest that marginal effect of past signal is close to zero or negative, whereas 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 future signal are the opposite: higher during recession period rather than ordinary periods. Such an observation indicates households are more sensitive to signals about past during ordinary periods and put more weights on signals about future when economic condition gets worse. It is also important to note that it doesn't necessarily mean they are more pessimistic during bad times because negative or close-to-zero marginal effects do not mean worse expectation on economic condition, rather it means expectation is less responsive to the signal considered.

Such a finding is obviously at odds to least square learning or noisy information models. In standard noisy information model with stationary Kalman Filters the marginal effect across time is fixed and depends only on variances of noise and prior. In least square learning models the marginal effect of signals may be time-varying when the data is limited but will be stabilized as more data is available to the learner. However the finding here suggests there's strong cyclicality of weights households put on specific signals, it is more consistent with the case that agents shift their attention to signals about future thus becoming more "forward-looking" during bad times of economy.

Moreover, such a finding doesn't only exist in expectation and signals on economic condition Δy, it also qualitatively holds for expectation and signals on unemployment status Δu. However there's still a caveat to the evidence I presented in this section: are the differences between marginal effects during ordinary and bad times statistically significant? As I discussed in **Section 3.2**, the naive estimates for marginal effect derived from functional estimations may be biased. To correct the potential biases and obtain estimates on average marginal effects with valid inference, I follow Chernozhukov *et al.* (2018) and obtain the DML Estimator. Then I can perform statistical tests on the difference between marginal effects in bad and ordinary times.

4.2.3 DML Estimator of Average Marginal Effects

To test whether the difference of marginal effects between ordinary and bad periods is statistically significant. I compute the DML Estimator following the procedures described in Section 3.2. Table 3 reports the estimated average marginal effect of past and future signals on expected economic condition 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 in **Section 4.2.2**.[26] And β_{ord} denotes the average marginal effect in periods other than the recession episodes. I then perform a Wald-test on $\beta_{rec} = \beta_{ord}$, the p-value is also reported in the table.

The first thing to notice in Table 3 is that the estimates are consistent with findings in Figure 4, where I use the naive estimate (as in equation (7)) of marginal effect computed at each point time. The signals about unemployment rate and inflation have negative impact on households' expectation about economic condition whereas signals about real GDP growth and interest rate usually have positive impact. The opposite is true for expectation on unemployment rate change. Both signals on past/current economic indicator and future are significantly affecting expectations. This suggests households are not complete adaptive learner and they have access to some information about future[27]. Moreover, signals on unemployment and real GDP growth have bigger impact when comparing to signals about interest rate and inflation.

The key message from Table 3 can be seen by comparing the marginal effects of the same signal between bad and ordinary period. For future signals on unemployment and real GDP growth, their marginal effects always have a bigger magnitude during bad episode, whereas

[26] For the same table using NBER recession dates for "bad periods", refer to **Appendix C.2**.

[27] This is consistent with findings from Barsky and Sims (2012).

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	$F_t\Delta u_{t+1}$	**−0.039*****	0.009**	<0.01	**0.029*****	0.007***	<0.01
		(0.005)	(0.003)		(0.004)	(0.002)	
	$F_t\Delta y_{t+1}$	**0.048*****	0.016***	<0.01	**−0.022*****	−0.009***	<0.01
		(0.005)	(0.003)		(0.002)	(0.001)	
	$F_t\Delta r_{t+1}$	0.025***	0.025***	0.93	−0.021***	−0.02***	0.89
		(0.007)	(0.003)		(0.005)	(0.002)	
	$F_t\pi_{t+1}$	0.015***	0.003**	<0.01	−0.008***	0.000	<0.01
		(0.003)	(0.001)		(0.002)	(0.001)	
Past Signal	Δu_t	−0.005	**−0.019*****	0.04	0.004	**0.010*****	0.12
		(0.005)	(0.004)		(0.004)	(0.002)	
	Δy_t	0.004*	**0.017*****	<0.01	−0.006***	**−0.01*****	0.04
		(0.002)	(0.002)		(0.001)	(0.001)	
	Δr_t	0.001	0.003***	0.44	−0.0003	−0.001	0.75
		(0.002)	(0.001)		(0.002)	(0.001)	
	π_t	−0.007***	−0.008***	0.86	0.004*	0.004***	0.93
		(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 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 effects of past signals are always bigger in ordinary episodes. The p-values on Wald test with null hypothesis: $H_0 : \beta_{bad} = \beta_{ord}$ range from 0.12 to less than 0.01 for these signals, which suggest the difference of marginal effects is statistically significant. However the same pattern does not hold true for signals on inflation and interest rate, with an exception of future signal on inflation.In fact average marginal effects on these signals are either insignificant or with small magnitudes. The marginal effect of signal can be interpreted as weights put on these signals when forming expectation. This pattern then indicates households shift their attentions from signals about past and current economic conditions to information about future economic conditions. Moreover, such an attention-shift is statistically significant. In other words, they are more adaptive learners when economic conditions are stable and become more forward-looking when situation gets worse.[28]

4.2.4 Decomposing Time-varying Marginal Effect

Now I have shown that households put more weights on signal from professional forecasters in bad times, meanwhile they rely less on realized macroeconomic statistics. However the explanation for such 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 in four categories: signals about economic conditions, signals about inflation, signals about interest rate and measure of news exposure about economic conditions.

As estimated ASF is non-linear, a proper way for variance decomposition is to use 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 expectation. Rather it can be interpreted as relative importance of these signals to explain the time-variation of marginal effects that I presented in **Section 4.2.2** and **Section 4.2.3**.

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 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 for time-variation observed, they explain up to 57% of the variation for marginal effect of past signal and 52% for that of future signal. News exposure on economic condition also play an important role especially for marginal effect of future signal, it explains 28% for future signal and 15% for past signal. With signals and

[28]Notice the time-varying marginal effect is created endogenously by the estimated model. One may wonder whether it's a repetition of the non-linear conditional expectation function. Intuitively the input signals such as realized unemployment rate are on average higher during recession. This means the average slope evaluated at those points are with higher absolute values. These concerns will be addressed in **Section 4.2.4**

[29]For same decomposition exercise of unemployment expectations refer to **Appendix C.3**

news exposure on economic condition alone I can explain as much as 72% and 80% of the total time-variation for marginal effects of past and future signal.

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 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 attentions put on signals about economic conditions and thus affecting the sensitivity of 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 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 play the most important role, which is around 70% of the total variation explained by these signals. Meanwhile state also accounts for a significant

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

share of the time-variation. It explains 17% and 13% respectively for marginal effects of past and future signals. This means the weight households put on economic condition signals depend on not only its current level, but also the state they accumulated from observing these signals in the past.

From the variance decomposition, I conclude types of signals that explain most of the time-variation of marginal effects are those about economic conditions and news exposure on these topics. However variance decomposition alone doesn't offer information about how these signals change marginal effects along time. It is possible that despite these two types of signals explain the most variation, they do not create the weight increase for future signal and decrease for past signal during bad times. To complete the picture, I present the time-varying marginal effects with only signals on economic conditions in Figure 5 and compare it with the actual marginal effects.

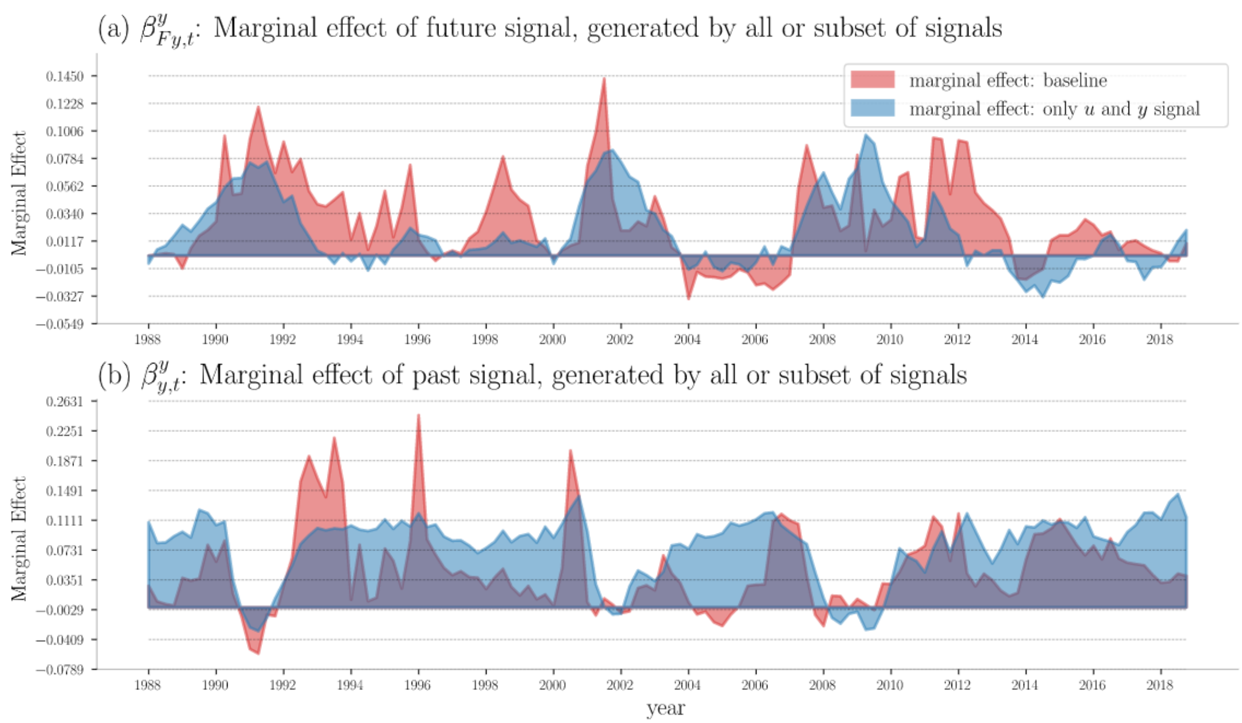


Figure 5: Time-varying marginal effect of past and future signal on real GDP growth. Top panel: marginal effect of future signal, $\beta^{Ey}_{Fy,t}$; bottom panel: marginal effect of future signal, $\beta^{Ey}_{y,t}$. The red curve: marginal effect created by estimated ASF with all signals. The blue curve: marginal effect created by ASF with only economic condition signals.

In Figure 5, the red curves are the time-varying marginal effects from estimated ASF with all signals as input. They are identical to those presented in **Section 4.2.2**. The blue curves are marginal effects computed from ASF using only actual economic condition signals as input[31], which are the same series I use to perform variance decomposition in Table 4. This figure shows strong evidence that economic condition signals generate the weight increase on future signals as well as drop of weight on past signals during bad times. They are indeed key driving forces for agents to change from adaptive learners to forward-looking when economic condition gets worse.

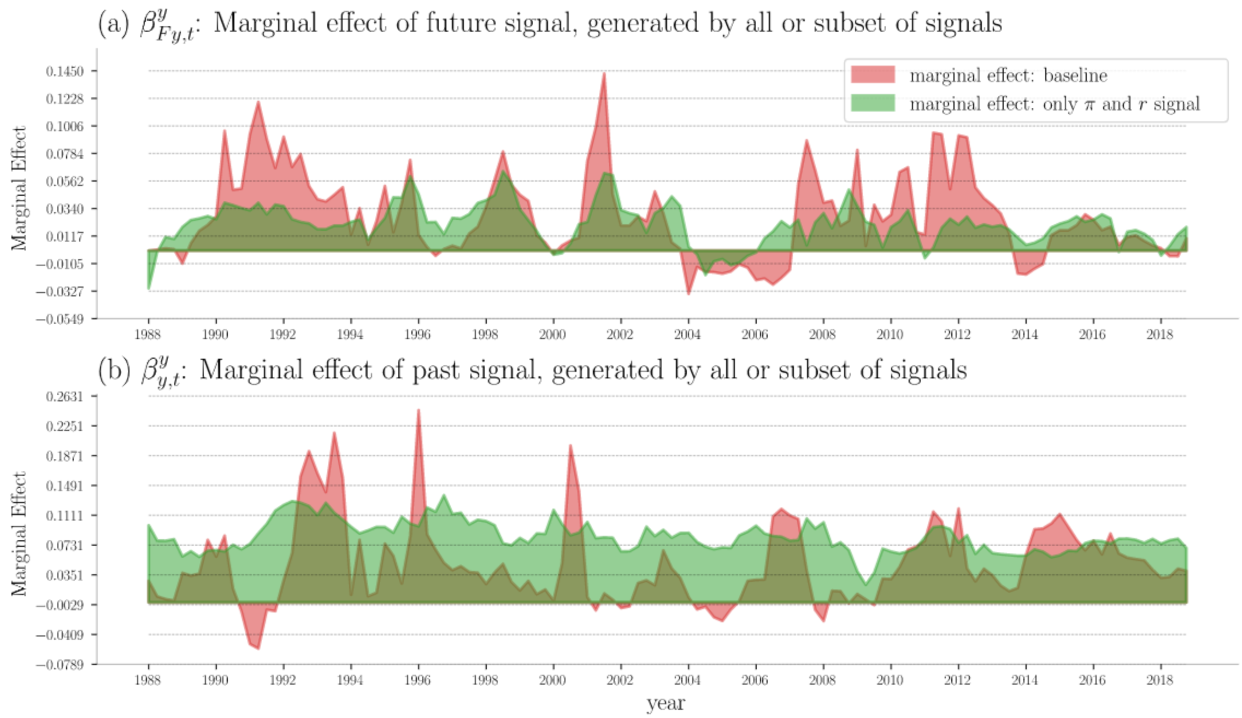


Figure 6: Time-varying marginal effect of past and future signal on real GDP growth. Top panel: marginal effect of future signal, $\beta^{Ey}_{Fy,t}$; bottom panel: marginal effect of future signal, $\beta^{Ey}_{y,t}$. The red curve: marginal effect created by estimated ASF with all signals. The blue curve: marginal effect created by ASF with only interest rate and inflation signals.

For comparison purpose, I also plot the marginal effects with only inflation and interest rate signals as input. These results are included in Figure 6. It is clear that with only inflation and interest rate signals, the ASF cannot generate the attention-shift we see in Figure 5. This also indicates that although inflation signal explain for 20% of time-variation in $\beta^{Ey}_{y,t}$ according to Table 4, it is not a primary driving force for the attention-shift which is the key finding I'm trying to explain.

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

4.3 Discussion

Applying the RNN approach to survey data on expectation for US consumers, I find the average expectation formation model for these households are non-linear and asymmetric, and they rely on signals about past during ordinary times but put more weights to signals about future economic conditions when the current economic status gets worse. With further exploration using the estimated ASF, I find the main reasons for such attention-shift are economic condition signals themselves together with level of news exposure on these topics.

For these findings to be consistent with FIRE, it has to be the actual fundamental law of motion that agent tries to learn is highly non-linear and state-dependent. For example, the contribution to unemployment from its past level has to be counter cyclical so that a rational agent with full information will appear to put lower weight on it during recessions. Such a law of motion is inconsistent with most of macroeconomic models.

These findings are clearly at odds with the standard noisy information model, which is typically assuming linear rules for expectation formation with implication of fixed marginal effect of signals across time. They are also hard to be reconciled with least square learning models as usually the time-variation of marginal effect appears when there are few data points to learn about the underlying structure, and it fades away when more information available to the learner.

One 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 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]. Figure 7 presents how news exposure affect 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 are 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 signal, whereas economic conditions alone explain more than 50%. Furthermore economic condition signals without news exposure successfully recreate the key attention-shift pattern. This

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

indicates economic condition signals explain a much bigger fraction of the time-variation.

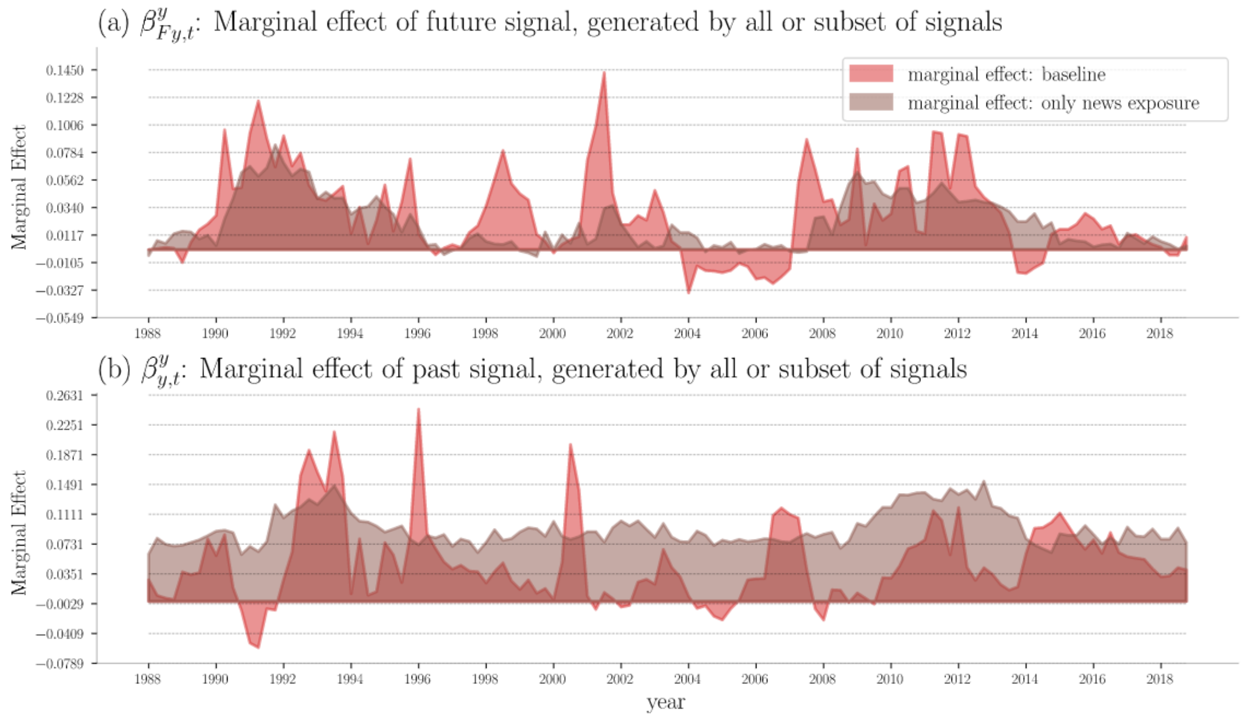


Figure 7: Time-varying marginal effect of past and future signal on real GDP growth. Top panel: marginal effect of future signal, $\beta^{Ey}_{Fy,t}$; bottom panel: marginal effect of future signal, $\beta^{Ey}_{y,t}$. The red curve: marginal effect created by estimated ASF with all signals. The blue curve: marginal effect created by ASF with only exposure of economic condition news.

Such a finding is then more in line with a rational inattention explanation for the phenomenon, in which agents will actively change the weights or attentions they put on past or future signals. In particular, households will put more efforts on searching for information about future during bad times. In rational inattention models, agents choose the allocation of attention optimally. For them to shift attention from past signal to future signal, it has to be the case that information about future is more valuable for them during bad times. This then helps to explain the non-linearity and asymmetry of the ASF shown in **Section 4.2.1**: when value of information changes as their perceived states of economy change, they will change the attention put on signals on realized economic statistics. In next section, I propose a simple rational inattention model to rationalize these properties found by RNN.

5 Model with Rational Inattention

So far I have documented brand news stylized facts about households' expectation formation model, using an innovative and more flexible semi-parametric approach with RNN. These new facts can be summarized with three key patterns: (1) households' expected economic condition is a non-linear and asymmetric function of signals about current economic status. In particular the shape of function is described in Figure 2. (2) When forming expectation on economic condition, agents are adaptive learners in ordinary times and become forward looking when state of economy gets worse, this is referred to as attention-shift. (3) The two major driving forces for the attention-shift are exposure to news about future economic status as well as signals on economic condition.

In this section I develop a simple two-period model where agents need to first allocate attention to signals about current and future state of economy and then make consumption and saving choices according to their beliefs on future economic condition. I will show the three stylized facts I documented in this paper is consistent with and can be explained by such a simple rational inattention model.

5.1 Household's Problem

There is a representative household that faces an individual consumption-saving problem. The household lives for two periods and get 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 $t+1$. At time t household will choose both consumption and saving with this risky asset, without knowing the value of d_{t+1} that is going to realize in $t+1$. At time $t+1$ the risky asset pays off and household consumes his total income in that period.

As the primary goal of this model is to assess agent's optimal choice of information structure to forecast variable at $t+1$, for simplification I assume the endowments $\{e_t, e_{t+1}\}$ are deterministic and the only uncertainty that agent faces comes from d_{t+1}. I then interpret d_{t+1} as the fundamental about economic condition in the future, as it accounts for all the uncertainty about agent's future income. If one consider saving as capital investment, with full depreciation d_{t+1} can be thought as productivity shocks in standard AK model.

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, agent forms belief on the return of risky asset and chooses consumption and saving according to this belief. Such a procedure in making optimal decisions is common in models with limited information. In rational inattention models, agents can choose the accuracy (variance) of signals he observes with an information cost. Signals with high accuracy and low variance will have high costs. I will discuss details about information structure and cost in **Section 5.2**. For now I will denote the information structure chosen optimally by agent as $\mathcal{I}_t$.

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

$$\begin{aligned} \max_{c_t, s_{t+1}} \quad & \mathbb{E}[u(c_t) + \beta u(c_{t+1}) | \mathcal{I}_t] \\ s.t. \quad & c_t + s_{t+1} = e_t \\ & c_{t+1} = (1 + d_{t+1}) s_{t+1} + e_{t+1} \end{aligned} \tag{14}$$

5.2 Information Structure

For agents to make 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 (15).

$$d_{t+1} = \rho d_t + \psi_{t+1} \tag{15}$$

To reflect the fact that there are 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.

$$\psi_{t+1} = \eta_t + \epsilon_{1,t+1} \tag{16}$$

$$\eta_t = \rho_\eta \eta_{t-1} + \epsilon_{2,t} \tag{17}$$

Both $\epsilon_{1,t+1}$ and $\epsilon_{2,t}$ are i.i.d and mean-zero shocks that follow normal distribution.

$$\begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix} \sim N(\mathbf{0}, \boldsymbol{Q}) \quad \boldsymbol{Q} \equiv \begin{bmatrix} \sigma^2_{1,\epsilon} & 0 \\ 0 & \sigma^2_{2,\epsilon} \end{bmatrix}$$

Such a formulation is similar to Barsky and Sims (2012) and it can be interpreted as "news shocks" described in Beaudry and Portier (2014). In general, this information may come from stock market, news or professionals. In this model for simplicity I consider that this information is contained in professional forecast. Throughout the model I will assume agent knows the correct law of motion of the stochastic return. In other words household is rational and the only reason for not forming perfect prediction on d_{t+1} is limitation of information he can get.

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 fundamental states by $\boldsymbol{X}_t \equiv \begin{bmatrix} d_t \\ \eta_t \end{bmatrix}$ and its prior as:

$$\boldsymbol{X}_0 \equiv \begin{bmatrix} d_0 \\ \eta_0 \end{bmatrix} \sim N(\mathbf{0}, \Sigma_0)$$

Then they face different information sets that contain various signals about the fundamentals. Some of these information may be passive, which means that agents will be exposed to them without cost. And some of these may be costly to acquire.

Signals agents are passively exposed to are on current state d_t. This is summarized as a Gaussian noisy signal $z_0 = d_t + \xi_0$, where $\xi_0 \sim N(0, \sigma_z^2)$. Such a signal can be thought as information agent picks up passively during daily life. For example, agents may get a rough idea about current economic condition when seeing friends or themselves getting unemployed or wage raise. These are information they are exposed to without putting in effort to collect thus incurring no cost. And this type of information may also be quite inaccurate, with a relatively high variance on noise.

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

$$F_t d_{t+1} = \rho d_t + \eta_t \tag{18}$$

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:

$$\begin{bmatrix} z_t^{spf} \\ z_t \end{bmatrix} \equiv \boldsymbol{Z}_t = G\boldsymbol{X}_t + \boldsymbol{\xi}_t \tag{19}$$

Agent is Bayesian Learner and form 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 noise is σ_z^2. Agent then updates his 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 d_t, unconditional prior mean and random noise in z_0. However agent has no control of variance of this noise σ_z^2. It will not be in agent's choice set and will be treated as given when agent solves the rational inattention problem later.

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

$$\hat{\boldsymbol{X}}_{t+1|t} \equiv \mathbb{E}[d_{t+1}|\boldsymbol{Z}_t] = A\{(I - KG)\hat{\boldsymbol{X}}_{t|0} + K\boldsymbol{Z}_{\boldsymbol{t}}\} \tag{20}$$

$$\Sigma_{t+1|t} = A\Sigma_{t|0}A' - AKG\Sigma_{t|0}A' + Q \tag{21}$$

And Kalman Gain is given by (22), where matrices A and G are given by exogenous parameters $\{\rho, \rho_\eta\}$ about the fundamentals.

$$K = \Sigma_{t|0}G'(G\Sigma_{t|0}G' + R)^{-1} \tag{22}$$

From (20)-(22), 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 agent will put higher weights on these signals. Each different choice of signal accuracy (represented by variance-covariance matrix on noise, R) give agent a different information set. Given different information set, agent will form different posterior belief even if the signals realized are the same. At this point, it's 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 $\epsilon_{1,t+1}$, agent with this information set can form a posterior belief on the distribution of d_{t+1} with mean being expressed as (18). This can be thought as 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 arbitrary variance-covariance matrix on noise, R, can be thought as in the middle of two information sets described above. For each information set $\mathcal{I}_t$ given, the agent will solve his optimization problem (14) accordingly. Different information sets will then result in different choices thus giving agent different expected utility. In this sense information has a value that can be evaluated with their expected utility. The rational inattention models suggest agent realizes the value of information and he chooses an optimal information set with a cost. Before I proceed to the full rational inattention set up, I will first illustrate the value of information in this model using different information sets described here.

5.3 Value of Information

In this section I explicitly compute 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 economy, d_t.

The agent solves his optimal consumption and saving choice given the information set $\mathcal{I}_t$. Defining for ease of notation $r_{t+1} = 1 + d_{t+1}$ and following (14), the maximization can be rewritten as:

$$\max_{s_{t+1}} \quad \mathbb{E}[u(e_t - s_{t+1}) + \beta u(r_{t+1}s_{t+1} + e_{t+1})|\mathcal{I}_t] \tag{23}$$

One key distinction I want to make between this model and standard rational inattention model is that I won't take the second-order approximation of the problem (23). Although this problem is well approximated in small deviations from steady state and the approach is commonly used in the rational inattention literature, it also leads to the result that only variance of posterior belief on state variable matters when agent chooses optimal signals. Such a result is intuitive when considering choices around the steady state, however it is not appropriate in my paper as I'm focusing explicitly on difference of attention choices during large deviations from the steady state. An important result my model will deliver is that mean of belief on economic state matters for agents' attention choices, this is the key mechanism to generate attention-shift that is found in empirical part.

To illustrate the value of information, I solve problem (23) under the three different information sets and evaluate agent's expected utility. For simplicity, I use a linear-quadratic utility function.

$$u(c_t) = c_t - bc_t^2$$

Such a function form also makes the point that the common mean-independent result of 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 utility function form. Given different information set $\mathcal{I}_t$, the first order condition for problem (23) then takes the form:

$$s^*_{t+1}(\mathcal{I}_t) = \frac{-1 + 2be_t + (\beta - 2b\beta e_{t+1}\mathbb{E}[r_{t+1}|\mathcal{I}_t])}{2b(1 + \beta\mathbb{E}[r^2_{t+1}|\mathcal{I}_t])} \tag{24}$$

From (24) it's clear that agent's optimal saving plan depends on his posterior belief on the return of risky asset, conditional on information set $\mathcal{I}_t$. Furthermore, it relies on both the conditional mean and the conditional variance of the return and is a non-linear function of them. It immediately follows that different information set will lead to differences in utility. When agent considers the value of information, they will evaluate the expected utility for each given information set $\mathcal{I}_t$.

I then numerically solve the model with 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 recursive Kalman Filter. The standard deviation of noise on passive signal is $\sigma_z = 0.22$. In Figure 8 I plot the 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 is available to agent. The thick blue curve is expected utility when d_t fully observable and the thick red curve is when both SPF and d_t are fully observable (the FIRE benchmark)[33]. The curves between the thick lines depicts the increase in expected utilities as precision of signal increases (or variance of noise decreases).

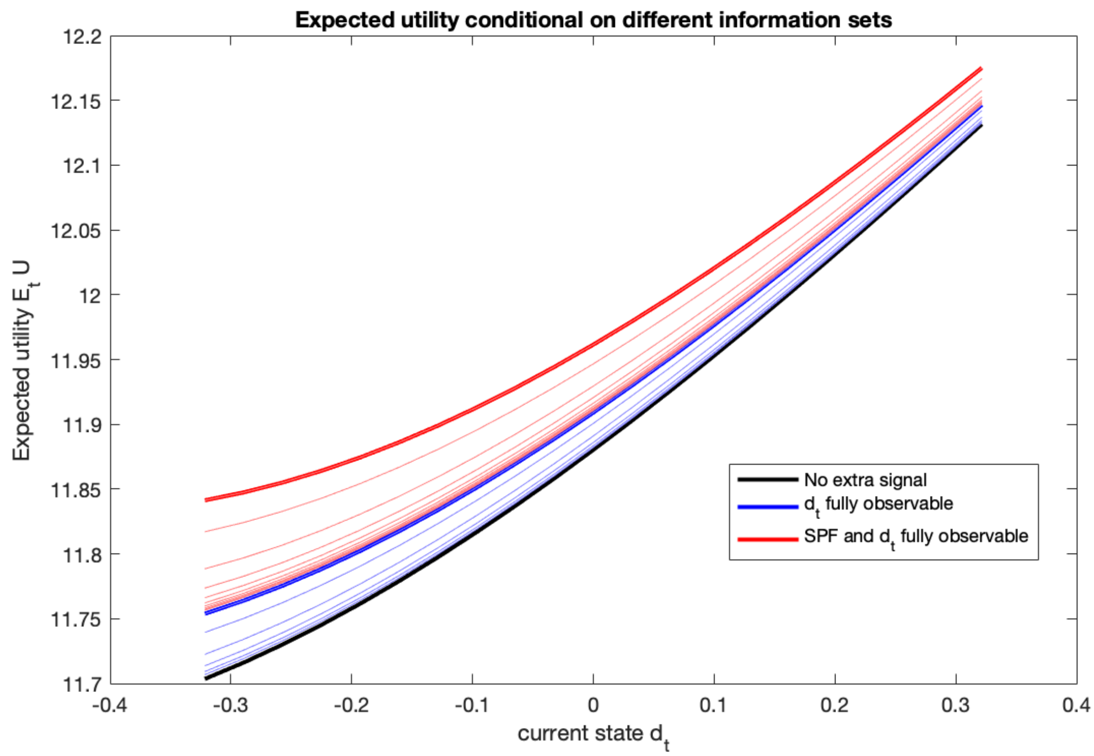


Figure 8: Black thick line: Expected utility when no more information other than initial passive signal on d_t; blue thick line: expected utility when d_t becomes fully observable; red thick line: when both SPF and d_t fully observable. Blue thin lines are expected utilities when there are noise attached 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 8. First more information improves agent's expected utility progressively: with more accurate signal on d_t, agent resolves the uncertainty about current state and his utility increases at any given d_t from black line to blue line; and it continues to increase as signal on SPF becomes more accurate, from blue curve to red

[33]With the specific law of motion assumed in (15) - (17) together with definition of SPF (18), the case with only SPF fully observable will coincide with the FIRE case.

curve. This is typical result from 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 same amount of information will increase agent's expected utility by more than the case when d_t is high. In other words, information is more valuable when 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 optimization problem.

The difference between expected utility comes from differences of optimal investment (24). 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 quadratic utility function, and re-write the expected utility as the following form:

$$\mathbb{E}[U(s^*_{t+1}(\mathcal{I}_t))] = -\mathbb{E}[\chi(s^*_{t+1}(\mathcal{I}_t) - \bar{s}_{t+1})^2] + M \tag{25}$$

Where $\chi = b(1+\beta r^2_{t+1})$ and $M = \mathbb{E}[\frac{(-1+2be_t+\beta r_{t+1}-2\beta b r_{t+1}e_{t+1})^2}{4b(1+\beta r^2_{t+1})}] + e_t - be^2_t + \beta e_{t+1} - \beta b e^2_{t+1}$[34]. The variable $\bar{s}_{t+1}$ is given by (26). It stands for the optimal invest under perfect foresight, when agent observes d_{t+1} perfectly.

$$\bar{s}_{t+1} = \frac{-1 + 2be_t + \beta r_{t+1} - 2b\beta r_{t+1}e_{t+1}}{2b(1+\beta r^2_{t+1})} \tag{26}$$

The transformed utility function (25) is usually referred 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[35]. From (25) it is obvious that if optimal choice of s is linear in state r_{t+1}, the expected utility only depends on posterior variance of r_{t+1} given information set $\mathcal{I}_t$. It is not related to posterior mean of states or realized state at time t.

Using the transformed expected utility (25), I can explore reasons for value of information decreasing in d_t. To see this, consider the cases with or without full information from

[34] The derivation refer to **Appendix D.1**

[35] 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}.

SPF. Conditional on 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 agent not having information from SPF can then be evaluated by differences between optimal savings with or without SPF observable, weighted by agent's subjective belief.

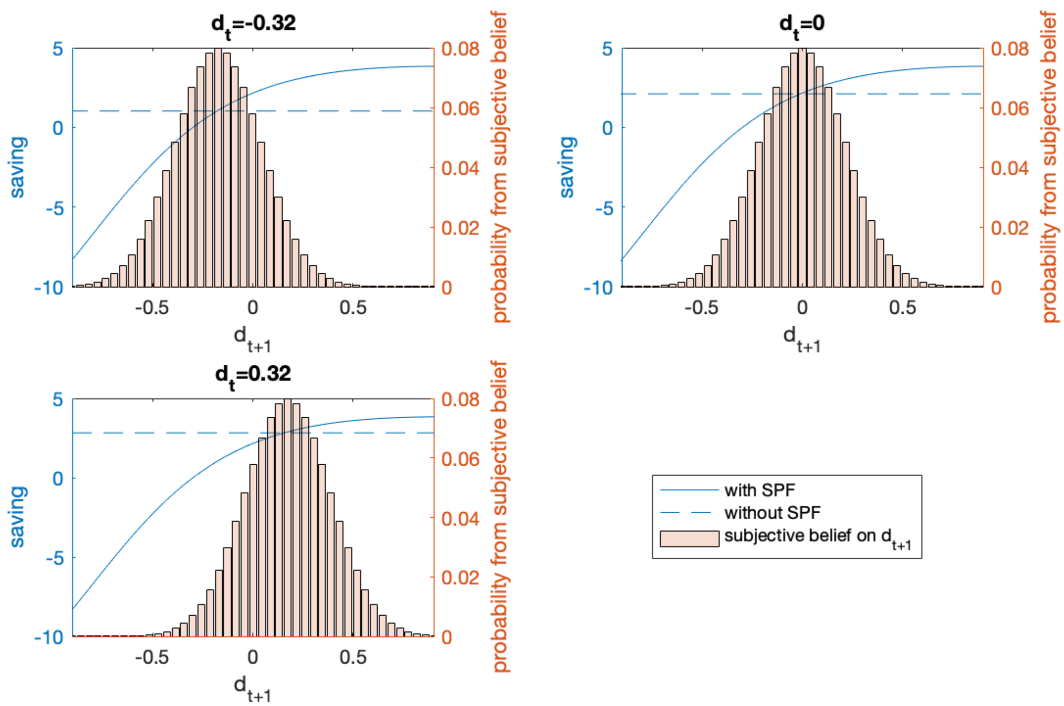


Figure 9: Solid line: optimal saving choice under full information: when both d_t and SPF are fully observable. Dash line: optimal policy when SPF signal is not available. Bar plot: agent's subjective belief on future state d_{t+1}, when SPF is not observable. Top left panel is when current state is very bad ($d_t = -0.32$), top right panel is when $d_t = 0$ and bottom left panel is when current state very good, $d_t = 0.32$.

In Figure 9 I depict the optimal saving choices at three realized values of d_t: when current state is bad ($d_t = -0.32$), neutral ($d_t = 0$) and good ($d_t = 0.32$). In each case I plot optimal saving choice as a function of future state d_{t+1}. The dotted line is the optimal saving that agent chooses when he only observes initial signal on d_t. It is a flat line because agent's choice does not depend on d_{t+1} (or realization of η_t) when SPF is not observable. The solid line is optimal saving choice when SPF is observable to 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 substitution effect becomes weaker as return on asset increases and finally dominated by income effect[36].

Now for agent without information from SPF, the solid line is not feasible. For a given realized d_t, the agent will evaluate his utility loss of not having information on η_t following (25). This is done by measuring the distance between optimal saving choices with and without information from SPF and compute the expected value of (the square of) this distance using their posterior belief on d_{t+1} (η_t). In Figure 9 this belief is shown with 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: value of information decreases in state of economy. Agent is willing to pay higher costs to acquire information as state of economy gets worse. This gives the key mechanism to create the time-varying marginal effect and non-linearity I documented with RNN. Because when agent 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. This is achieved with the full rational inattention model in next section.

5.4 Model with Rational Inattention

In the model with Rational Inattention, given information set, agent will face the same two period consumption-saving problem as in **Section 5.1**. Agents face the same uncertainty generated by two fundamentals described by (15)-(17). They also know information has value as described in **Section 5.3**. The major difference here is that agent now can costly

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

choose the optimal signals on fundamentals that he can observe in his information set. At this point it's useful to describe the time-line for agent's problem.

Time-line of Agent's Problem: The agent problem can be described as following steps:

1. At beginning of time t, agent is exposed to a signal about current state d_t. The signal was denoted as z_0. Agent is endowed with some normal prior on fundamentals:

$$\boldsymbol{X}_0 \equiv \begin{bmatrix} d_0 \\ \eta_0 \end{bmatrix} \sim N(\mathbf{0}, \Sigma_0)$$

2. Agent chooses optimal signals in information set $\mathcal{I}_t$, knowing the consumption-saving choices he make later will depend on the information set he chooses. The form of signals and contents in information sets are same as those in **Section 5.2**.

3. Given the chosen information set, agent chooses consumption and saving to maximize his expected utility, conditional on signals observed.

Information Set and Cost: Agent's initial information set contains his prior on foundmentals, $\boldsymbol{X}_0$ as well as the initial passive signal on d_t: z_0. This is because agent has no control on z_0. This information set can then be fully summarized with updated prior about $\boldsymbol{X}_t$. Denote the initial information set as $\mathcal{I}_0 = \{\boldsymbol{X}_{t|0}\}$, for the entire problem agent will take this information set as given. The information set after agent choosing precision of signals can be then defined as:

$$\mathcal{I}_t = \mathcal{I}_0 \cup \{\boldsymbol{Z}_t\} \tag{27}$$

Where $\boldsymbol{Z}_t$ is vector of signals defined in (19).

Information comes with a cost. Following Sims (2003) I measure the cost of acquiring more information in set $\mathcal{I}_t$ with difference of 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 information cost can be represented by posterior variance-covariance matrices. Denoted as κ, equation (28) formally defines information cost.

$$\kappa = \mathcal{H}(d_{t+1}|\mathcal{I}_0) - \mathcal{H}(d_{t+1}|\mathcal{I}_t) = \frac{1}{2}log_2(\frac{det\Sigma_{t+1|0}}{det\Sigma_{t+1|t}}) \tag{28}$$

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 $\mathcal{I}_t$.

Agent's Optimization Problem: Agent's problem comes in two steps. First 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 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:

$$\max_{\sigma^2_{1,\xi},\sigma^2_{2,\xi}} \quad \mathbb{E}[u(e_t - s^*_{t+1}) + \beta u(r_{t+1}s^*_{t+1} + e_{t+1})|\mathcal{I}_0] - \lambda\kappa \tag{29}$$

$$s.t. \quad s^*_{t+1} = argmax_{s_{t+1}} \quad \mathbb{E}[u(e_t - s_{t+1}) + \beta u(r_{t+1}s_{t+1} + e_{t+1})|\mathcal{I}_t] \tag{30}$$

$$\kappa = \frac{1}{2}log_2(\frac{det\Sigma_{t+1|0}}{det\Sigma_{t+1|t}}) \tag{31}$$

Equation (29) describes the first step of agent's problem. He chooses optimal precision of signals knowing that his choice of saving depends on realization of these signals. In this expectation agent will take into account uncertainty that comes from both underlying true states and noise in realized signals. His choice on precision of signals will affect $\mathcal{I}_t$ in (30) as well as information constraint (31).

The agent's problem can then be solved backwards. The trade-off agent faces in solving this problem is then between the value of more information and its cost. As shown in Figure 8, lower $\sigma_{1,\xi}$ and $\sigma_{2,\xi}$ will increase expected utility. Meanwhile more accurate signals will also increase information cost κ, as accurate signals decrease the posterior variance of agent's belief. Because agent observes an initial signal z_0 which contains information about d_t. Then agent's optimal choice of signal precision will depend on d_t[37]: when d_t is negative, information becomes more valuable to agent thus they are willing to choose higher precision (lower variance) for signals.

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

5.5 Results

I solve the rational inattention problem (29)-(31) by simulation using the following parametrization:

Table 5: 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
$\sigma_{1,\epsilon}$	0.15	$\sigma_{2,\epsilon}$	0.15
σ_z	0.46	λ	0.15

The purpose of this model is to illustrate non-linear functional form and state-dependency weights can be generated with rational inattention. 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 agent's optimal choices of signal variance as well as the model implied weight on current (d_t) and future (SPF) signals. These results are included in Figure 10.

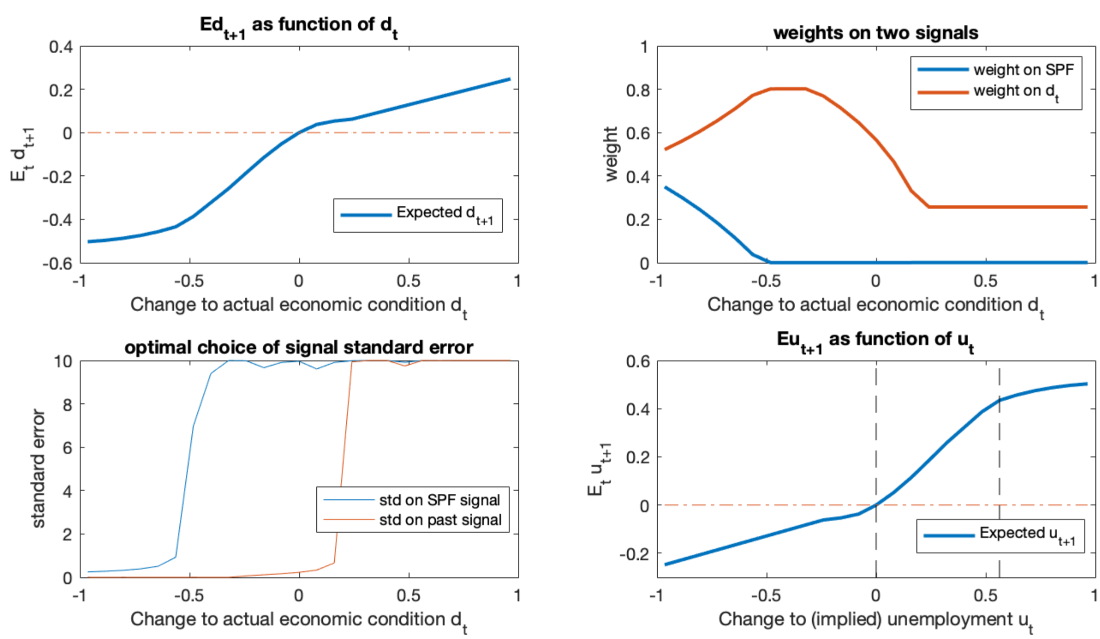


Figure 10:

Top left panel of Figure 10 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: agent chooses variance on both signals to be as high as 100, which imply almost 0 weights on both these new signals. 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 $-.05 < d_t < 0$, 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, 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 an 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.

6 Conclusion

How does household form expectation using a rich set of macroeconomic signals? This paper explores the answer to this question by proposing an innovative Generic Learning Model that is flexible in functional forms and time-dependency that describe the relationship between signals and expectational variables. The unknown functional form is estimated with a state-of-art Recurrent Neural Network. This method can recover any function forms considered by the Generic Learning Model, including those most commonly used in the learning literature, such as noisy information, least square learning and hidden markov models. Following the Double De-biased Machine Learning approach developed by Chernozhukov *et al.* (2018), I then obtain estimators on average marginal effect of signals with valid inferences.

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 condition is a non-linear and asymmetric function of signals on well-being of economy; (2) the attentions to past and future signals in the Generic Learning Model are highly state-dependent, agents are adaptive learner in ordinary periods and become forward looking as state of economy gets worse; (3) both signal and exposure to news on economic condition play important role in creating the attention-shift. These findings are at odds with noisy information models and least square learning models that are widely used in the literature. Agents' attention-shift from past to future signals comes from two distinct reasons: information about future flowing from professionals to households as firstly proposed by Carroll (2003) and agents actively choose to allocate their limited attention on information about future as economic situation deteriorates. These findings then suggest the dynamics of economic variables may be largely different in ordinary and recession episodes as the information contents in households expectation differ as state of economy changes. Another implication is that policy that features forward guidance and expectation management may be ineffective as economic conditions become stable as agents pay less attention to information about future.

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

References

Almosova, A. and Andresen, N. (2018). *Forecasting Inflation with Recurrent Neural Networks.*

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.

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.

Bordalo, P., Gennaioli, N. and Shleifer, A. (2018). Diagnostic expectations and credit cycles. *The Journal of Finance*, **73** (1), 199–227.

Caplin, A. (2016). Measuring and modeling attention. *Annual Review of Economics*, **8** (1), 379–403.

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.

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.

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

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. (2018). Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *ArXiv*, **abs/1809.09953**.

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

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.

Kuan, C.-M. and Liu, T. (1995). Forecasting exchange rates using feedforward and recurrent neural networks. *Journal of Applied Econometrics*, **10** (4), 347–64.

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.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, **1**, 19 – 46.

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

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

Nakamura, E. (2005). Inflation forecasting using a neural network. *Economics Letters*, **86** (3), 373 – 378.

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.

R., C. T. and Hall, A. S. (2017). Macroeconomic indicator forecasting with deep neural networks. *Federal Reserve Bank of Kansas City, Research Working Paper*, **17-11**.

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.

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 $\mathbb{E}[Y_{i,t+1|t}|\{Z_{i,\tau}\}_{\tau=0}^{t}]$. This conditional expectation function can then be written as:

$$\begin{aligned}
\mathbb{E}[Y_{i,t+1|t}|\{Z_{i,\tau}\}_{\tau=0}^{t}] &= \int F(\Theta_{i,t}) d\mathcal{F}_{\Theta_{i,t}}(\Theta_{i,t}|\{Z_{i,\tau}\}_{\tau=0}^{t}) \\
&= \int F(\Theta_{i,t}) \mathcal{P}_{\Theta_{i,t}}(\Theta_{i,t}|\{Z_{i,\tau}\}_{\tau=0}^{t}) d\Theta_{i,t} \\
&= \int \left(\int F(\Theta_{i,t}) \mathcal{P}_{\Theta_{i,t}}(\Theta_{i,t}|\{Z_{i,\tau}\}_{\tau=0}^{t}, \Theta_{i,t-1}) d\Theta_{i,t} \right) \mathcal{P}_{\Theta_{i,t-1}}(\Theta_{i,t-1}|\{Z_{i,\tau}\}_{\tau=0}^{t}) d\Theta_{i,t-1}
\end{aligned} \tag{32}$$

Where in equation (32), the first equality holds from Assumption 2, that the expectation formation process admits a latent variable structure, with a finite dimensional latent variable $\Theta_{i,t}$. The conditional CDF of variable X is represented by $\mathcal{F}_X$ and conditional PDF is represented by $\mathcal{P}_X$. The third equality then 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 $\epsilon_{i,t}$:

$$\begin{aligned}
\mathcal{P}_{\Theta_{i,t}}(\Theta_{i,t} = r'|\{Z_{i,\tau}\}_{\tau=0}^{t}, \Theta_{i,t-1}) &= \mathcal{P}_{\epsilon_{i,t}}(H(\Theta_{i,t-1}, Z_{i,t}, \epsilon_{i,t}) = r'|Z_{i,t}, \Theta_{i,t-1}) \\
&= \mathcal{P}_{\Theta_{i,t}}(\Theta_{i,t} = r'|Z_{i,t}, \Theta_{i,t-1})
\end{aligned} \tag{33}$$

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

$$
\begin{aligned}
\mathcal{P}_{\Theta_{i,t}}(\Theta_{i,t} = r' | Z_{i,t} = z, \Theta_{i,t-1} = r) &= \mathcal{P}_{\epsilon_{i,t}}(H(r, z, \epsilon_{i,t}) = r') \\
&= \mathcal{P}_{\epsilon_{i,t+s}}(H(r, z, \epsilon_{i,t+s}) = r') \\
&= \mathcal{P}_{\Theta_{i,t+s}}(\Theta_{i,t+s} = r' | Z_{i,t+s} = z, \Theta_{i,t+s-1} = r) \quad \forall s > 0
\end{aligned} \tag{34}
$$

In other words, from (33) and (34) the latent variable $\Theta_{i,t}$ is a time-homogenous Markov Process conditional on realization of $Z_{i,t}$. Now one can discretize the continuous-state Markov Process[38]. Denoting 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 $\{p_{r,r'}(z)\}$. Notice the transition probability will be a function of realized signal in each period z, so that for different realization of Z, we have a different Markov Chain. Now consider a finite dimensional variable:

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

Then it follows immediately from (32) 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.

And it is obvious that $\theta_{i,t}$ 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 $t = 0$:

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

For $t = 1$:

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

[38]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.

Where the second equality from above follows from Markov Property (33) then with time-homogeneity (34), one can get time t relation by induction:

$$\theta_{i,t}^{r'} = \sum_{r=1}^{N_r} p_{r,r'}(Z_{i,t})\theta_{i,t-1}^{r} \tag{35}$$

Equation (35) 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.

B.1 Neyman Orthogonal Moment Condition

1. Begin by declaring Joint Objective function, **at each time point** t, denote $X \equiv \{X_{i,\tau}\}_{\tau=0}^{t}$:

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

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

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:

$$\begin{aligned}\mathbb{E}[\ell(\{y, X, \theta_{-1}\}; \beta_t, g_t)] &= \int \mathbb{E}[\ell()|X, \theta_0]P(X, \theta_{-1})d(X, \theta_{-1}) \\ &\equiv \int \mathcal{L}(\{X, \theta_{-1}\}; \beta_t, g_t, g_{t,x})d(X, \theta_{-1})\end{aligned} \tag{36}$$

Using Euler-Lagrangian Equation:

$$\begin{aligned}0 &= \frac{\partial \mathcal{L}}{\partial g_t} - \sum_j^J \frac{\partial}{\partial x_t^j}(\frac{\partial \mathcal{L}}{\partial g_{t,x}^j}) \\ &= \underbrace{-(\mathbb{E}[y|X, \theta_{-1}] - g_t(X, \theta_{-1}))P(X, \theta_{-1})}_{\equiv \frac{\partial \mathcal{L}}{\partial g_t}} \\ &- \sum_j^J \frac{\partial}{\partial x_t^j} \underbrace{\Big(- (\beta_t^j - g_{t,x}^j(X, \theta_{-1}))P(X, \theta_{-1})\Big)}_{\equiv \frac{\partial \mathcal{L}}{\partial g_{t,x}^j}} \\ &= -(\mathbb{E}[y|X, \theta_{-1}] - g_t(X, \theta_{-1}))P(X, \theta_{-1}) + \\ &\sum_j^J \Big(- g_{t,xx}^j(X, \theta_{-1})P(X, \theta_{-1}) + \frac{\partial P(X, \theta_{-1})}{\partial x_t^j}(\beta_t^j - g_{t,x}^j(X, \theta_{-1}))\Big)\end{aligned} \tag{37}$$

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:

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

Plug this into the moment condition:

$$\mathbb{E}[\beta_t^j - f \circ h_{x^j}(\theta_{i,t-1}, X_{i,t}) + \frac{\partial ln(P(X, \theta_0))}{\partial x^j}(y - f \circ h(\theta_{i,t-1}, X_{i,t})] = 0$$

The second problem can be solved by assuming dependency of $X_{i,t}$ and $X_{i,t-s}$. As $\theta_{-1} = 0$ which is deterministic. The above equation leads to moment condition (9) given the fact that $g(\{Z_\tau\}_{\tau=0}^{t}, \theta_{-1}) \equiv f \circ h(\theta_{i,t-1}, X_{i,t})$

B.2 Verifying Moment Condition is Orthogonal

This can be done by computing the Frechet Derivative of the moment condition denoted as $\mathbb{E}[\psi(\beta, X, g)]$, for sake of simplified notation, I drop the t and consider 1 dimensional case, but the appliction can be easily extended to multidimensional case.

$$\psi(\beta, g, X) \equiv \beta - g'(X) + \frac{P'(X)}{P(X)}(\mathbb{E}[y|X] - g(X)) \tag{38}$$

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

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

The Frechet Derivative along direction v is given by:

$$
\begin{aligned}
F(g+v)-F(g) &= \mathbb{E}[-v'(X)-\frac{P'(X)}{P(X)}v(X)] \\
&= \lim_{\delta\to 0}\mathbb{E}[-\frac{v(X+\delta)-v(X)}{\delta}-\frac{P(X)-P(X-\delta)}{P(X)\delta}v(X)] \\
&= \lim_{\delta\to 0}1/\delta[-\int_X v(X+\delta)P(X)dX+\int v(X)P(X)dX \\
&- \int v(X)P(X)dX+\int v(X)P(X-\delta)dX] \\
&= \lim_{\delta\to 0}1/\delta[\int_y v(y+\delta)P(y)dy-\int_x v(x+\delta)P(x)dx] \\
&= 0
\end{aligned} \tag{39}
$$

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). To apply this theorem one needs to verify three condition[39]:

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

$$\psi(\beta,g,X)=\psi^a(X,g)\theta+\psi^b(X,g)$$

2. (Near) Neyman Orthogonality of score $\psi(\beta,g,X)$;

3. Fast enough convergence of nuisance parameter g. 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 $\mathbb{E}[(\hat{g}(X)-g_0(X))^2]^{1/2}=o(n^{-1/4})$.

Condition 1 is obvious given the Neyman Orthogonal score derived in Appendix B.1: equation (38) is linear in β. Condition 2 is verified in Appendix B.2. The convergence speed requirement in condition 3 is offered by Theorem 1 of Farrell *et al.* (2018). To achieve the convergence speed described there, one needs to put restrictions on width and depth of neural network used to approximate $g(.)$. My baseline architecture satisfies these restrictions.

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

B.4 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 (5) 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.

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:

$$\begin{bmatrix} \pi_t \\ L_t \end{bmatrix} \equiv X_t = \boldsymbol{A} X_{t-1} + \epsilon_t \tag{40}$$

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 (40). Agents do not observe X_t directly, instead they observe a noisy signals about it. Their observational equation is:

$$\begin{bmatrix} \pi_{i,t} \\ s_{i,t} \end{bmatrix} \equiv O_{i,t} = \boldsymbol{G} X_t + \nu_{i,t} \tag{41}$$

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

$$\begin{bmatrix}\pi_{i,t+1|t}\\ L_{i,t+1|t}\end{bmatrix} \equiv X_{i,t+1|t} = \boldsymbol{A}(X_{i,t|t-1} + \boldsymbol{K}(O_{i,t} - \boldsymbol{G}X_{i,t|t-1})) \tag{42}$$

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 (42) as equation (43) and (44), 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.

$$\hat{Y}_{i,t+1|t} = \boldsymbol{A}\Theta_{i,t} \tag{43}$$

$$\Theta_{i,t} = (\boldsymbol{A} - \boldsymbol{KGA})\Theta_{i,t-1} + \boldsymbol{KG}X_t + \boldsymbol{K}\nu_{i,t} \tag{44}$$

Average Structural Function I then turn to the ASF implied by Kalman Filter (43) and (44). 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:

$$\theta_{i,t} = \mathbb{E}[\Theta_{i,t}|\{X_\tau\}_{\tau=0}^t] \tag{45}$$

Taking the expectation of (43) and (44) conditional on history of the observable $\{X_\tau\}_{\tau=0}^t$ it immediately follows:

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

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 (42) 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[40]. 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 (42). 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[41]; (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 and expectational variable as well as obtain comparable estimate on signals

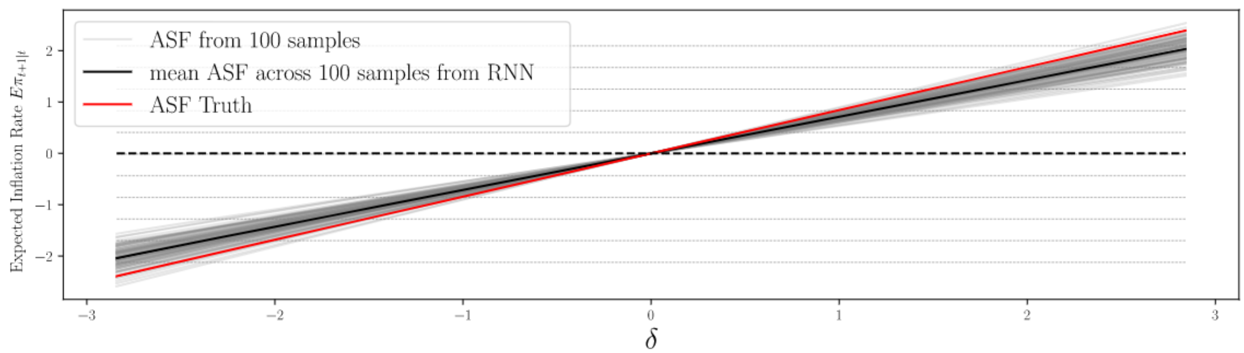
[40]For example, when agent form expectation on inflation, if they believe in a three equation New Keynesian Model, they may also want to infer demand and supply shocks as unobserved states. In a Kalman Filter that takes only inflation as unobserved state, OLS will suffer from omitted variable problem.

[41]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.

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 11. The red solid line is the true Average Structural Function implied by the Kalman Filter (42) 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 11 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.

(a) $E\pi_{t+1|t}$ when π_t change by δ



(b) $E\pi_{t+1|t}$ when $s_{i,t}$ change by δ

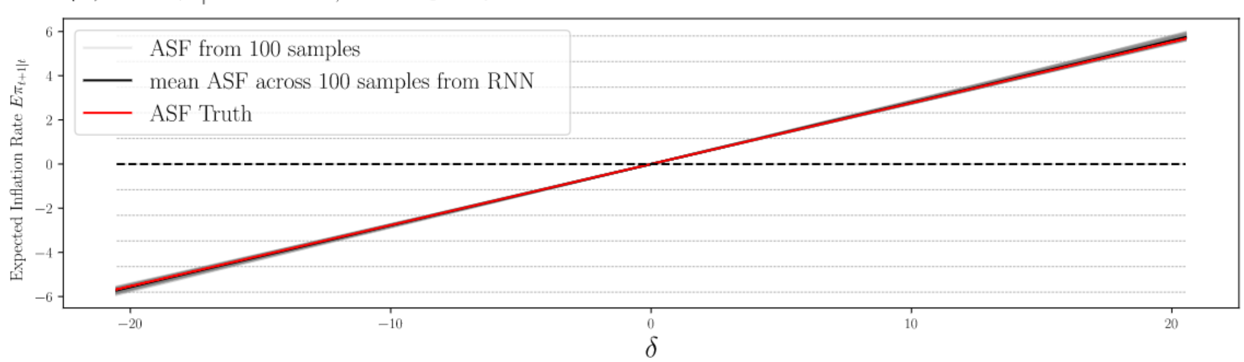


Figure 11: Estimated Average Structural Function from random samples using RNN. Function depicts change of expected variable in response to corresponding signal change by δ. Panel (a): expected inflation as function of inflation signal π_t. Panel (b): expected inflation as function of private signal $s_{i,t}$. 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 6: 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 Appendix on Empirical Findings

C.1 More on Time-varying Marginal Effect

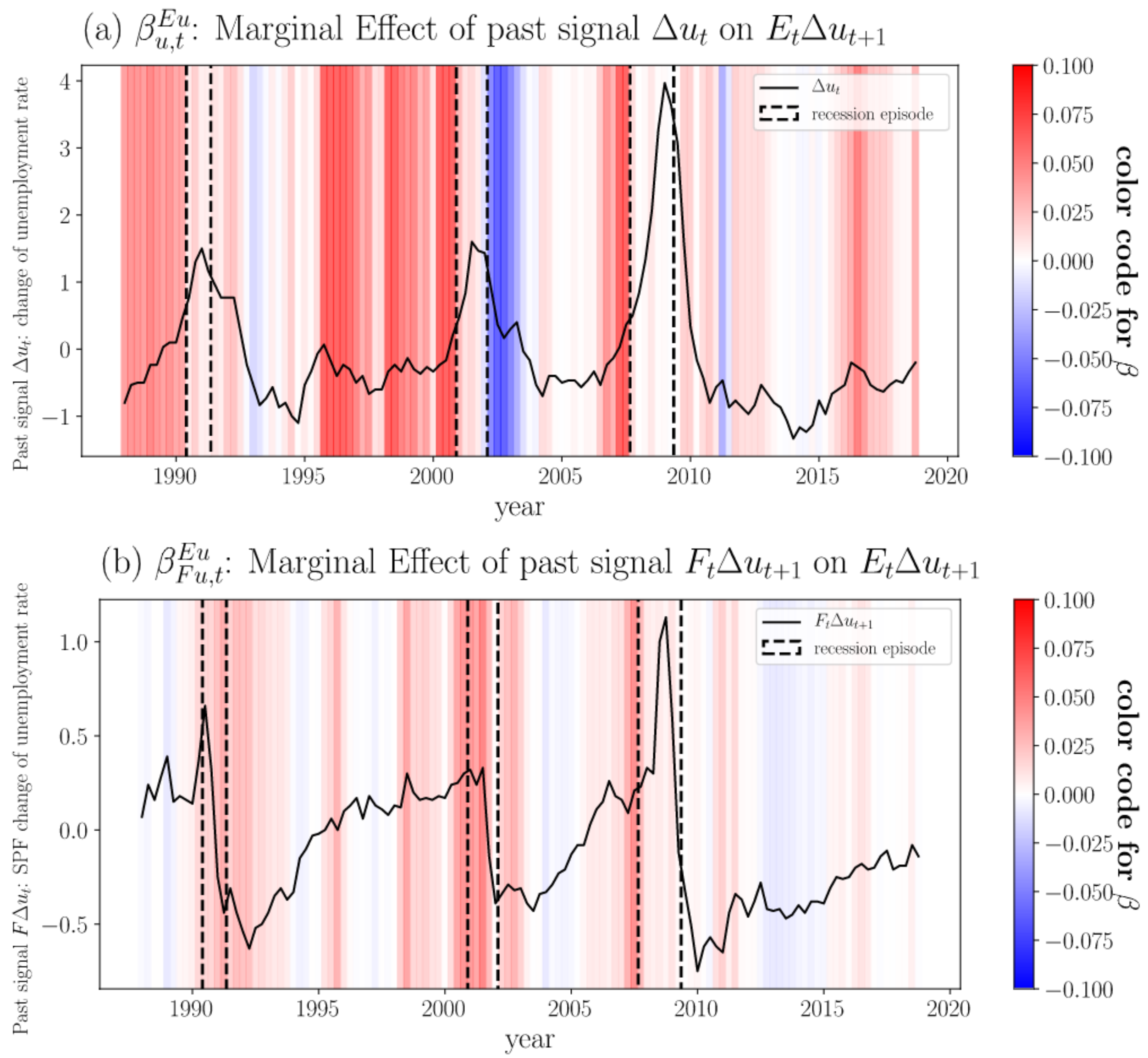


Figure 12: Color bars in panel (a): the marginal effects of unemployment signal Δu_t on expected unemployment rate increase next year $E\Delta u_{t+1|t}$. Panel (b): the marginal effects of professionals' forecasts about unemployment rate change next year $F\Delta u_{t+1|t}$ on expected unemployment rate change. Red color: positive marginal effect; blue color: negative marginal effect. Black solid line: data on frequency of news about recession.

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 12 below. It shows the same pattern holds as in Figure 4: 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 13 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.

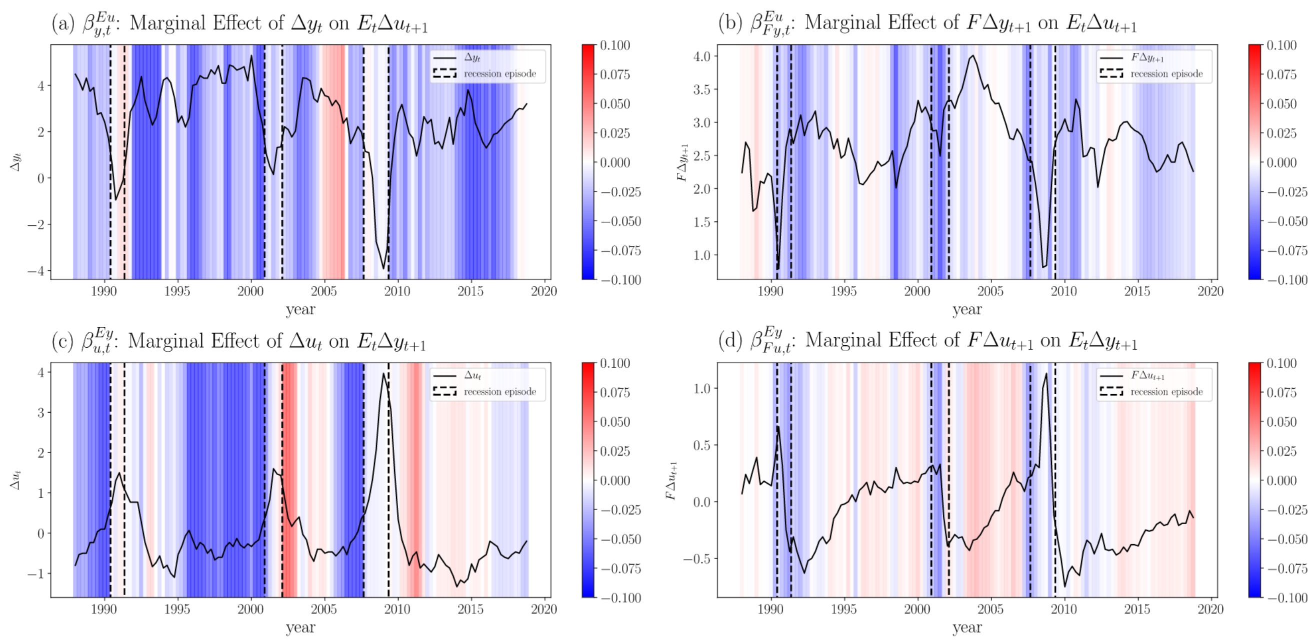


Figure 13:

However these attention shift during recession and ordinary period only holds significantly for expectations and signals related to indicators about economic conditions. Figure 14 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.

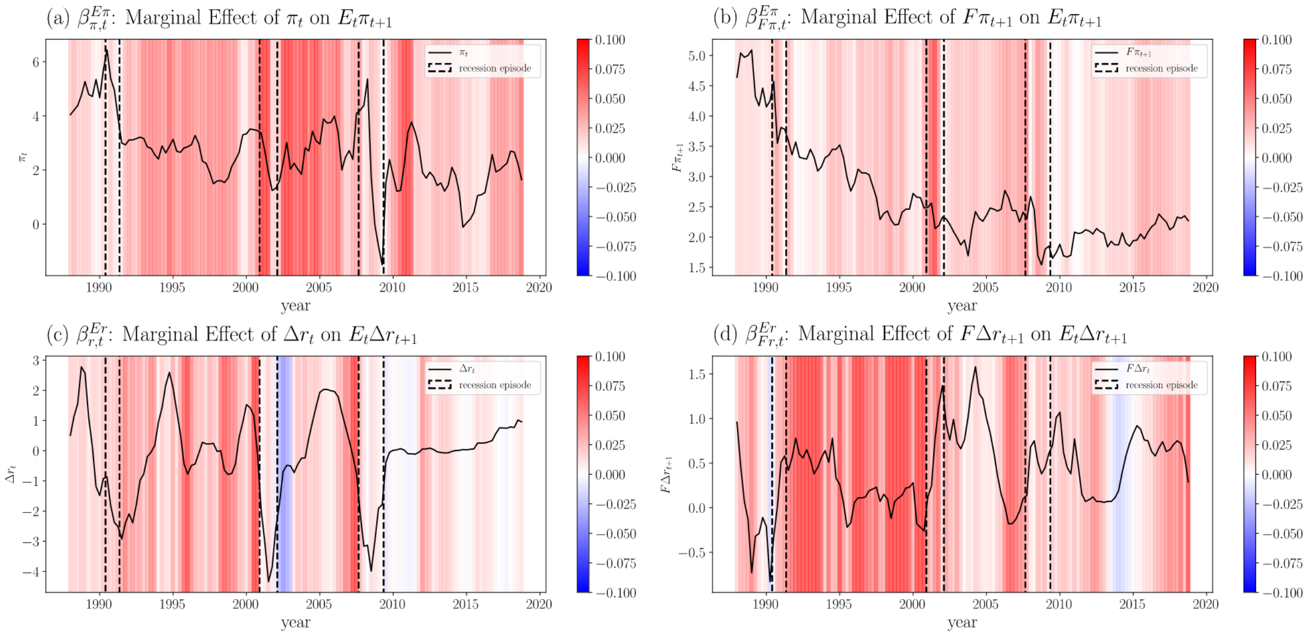


Figure 14:

C.2 Robustness of DML using NBER Recessions

Table 7: 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	$F_t\Delta u_{t+1}$	**−0.046*****	0.006**	<0.01	**0.03*****	0.008***	<0.01
		(0.006)	(0.003)		(0.004)	(0.002)	
	$F_t\Delta y_{t+1}$	**0.049*****	0.019***	<0.01	**−0.024*****	−0.01***	<0.01
		(0.006)	(0.003)		(0.003)	(0.001)	
Past Signal	Δu_t	−0.023***	−0.015***	0.3	0.015***	0.008***	0.12
		(0.007)	(0.003)		(0.004)	(0.002)	
	Δy_t	0.006*	**0.015*****	<0.01	−0.006***	**−0.01*****	0.1
		(0.003)	(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 7 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.

C.3 Variance Decomposition for Unemployment Expectation

In Table 8 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.

Table 8: 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%	

D Derivation in Rational Inattention Model

D.1 Quadratic Loss Function

Start from LHS of equation (25), to same notations I denote $s^*_{t+1}(\mathcal{I}_t)$ as s_{t+1}:

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

The last line gives the RHS of equation (25). Now notice, 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. However, the key difference here is s^*_{t+1} is non-linear in fundamentals. 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).