Directed attention and nonparametric learning

Ian Dew-Becker and Charles G. Nathanson*

March 19, 2017

Abstract

We study the optimal information and consumption choice of an ambiguity averse agent with uncertainty about income dynamics. The novelty of the paper is that agents choose what aspects of the income process to learn about. We solve the model in closed form and show that the utility-optimal information structure provides maximal precision about income dynamics at the very lowest frequencies, which have the greatest effect on utility. Deviations of consumption from the full-information rational expectations benchmark are then predicted to be largest at high frequencies, so the model can explain why consumption tracks predictable changes in income but is close to a random walk in the long-run. Furthermore, whereas ambiguity aversion typically leads agents to act as though shocks are more persistent than the truth, endogenous learning here eliminates that effect.

1 Introduction

A growing literature studies economic behavior in the face of model uncertainty, while at the same time there is a large amount of recent work that studies optimal allocation of attention.[1] Those two areas are obviously related: the economy is highly complex, so people are unlikely to be able to understand all of it, and they must choose how to allocate their limited attention and information processing abilities. Furthermore, information acquisition is not free, so we would not necessarily expect people to be perfectly informed about everything.

Surprisingly, though, there is little or no research that studies the implications of directed attention in the face of model uncertainty.[2] The contribution of this paper is to study the behavior of an agent who allocates attention across different aspects of a model. We show that optimal learning about model features has important and interesting implications for behavior. On the

*Dew-Becker: Northwestern University and NBER. Nathanson: Northwestern University. We appreciate helpful comments from Ben Hebert, Mikkel Plagborg-Møller, Konstantin Milbradt, and seminar participants at Northwestern.

[1] On model uncertainty, most revelant for us is recent work on consumption under model uncertainty, e.g. Hansen, Sargent, and Tallarini (1999), Wang (2004, 2009), Luo (2008), Luo and Young (2010), but there is a large broader literature. See Sims (2003), Veldkamp (2011), and many citations therein for work on directed attention.

[2] There is substantial past work on directed learning (e.g. Van Nieuwerburgh and Veldkamp (2006), Peng and Xiong (2006), Veldkamp (2006), and Barron and Ni (2008)), but we are not aware of work that examines the choice of what part of a dynamic process to learn about.

one hand, it naturally leads to excess sensitivity of consumption to high-frequency components of income, as observed empirically. At the same time, though, we show that optimally directed attention drives the consumption policy closer to the optimum at lower frequencies than it would be if model uncertainty were purely exogenous.

More concretely, we study the problem of an ambiguity averse agent who is uncertain about the dynamics of an exogenous and untradeable income process. That is, the agent solves the standard minimax problem studied by Gilboa and Shmeidler (1989), making consumption decisions that are robust to model uncertainty. Directed attention, where our contribution lies, is then modeled by making the level and type of model uncertainty endogenous. When the agent obtains more information about the world, the accuracy of her approximate model increases and her concerns about ambiguity decrease.[3]

Rather than learning about a finite set of parameters, in an ARMA model, for example, we consider a setting where the agent faces nonparametric model uncertainty and has uncertainty about the entire spectral density of the income process (which completely characterizes its behavior here). In choosing information, the agent decides the precision of signals that she receives about the magnitude of the spectrum of income at each frequency. The choice of information is made to maximize utility in the minimax setting.

All three phases of the optimization are analytically solvable. Consumption, and hence realized utility, depends primarily on the agent's beliefs about the low-frequency characteristics of income. Low-frequency shocks that drive permanent income have the greatest impact on consumption volatility. When nature chooses an unfavorable model, then, it tends to select one with high power at low frequencies, implying that income is strongly persistent and consumption risky. In other words, people naturally fear highly persistent income processes.

That force is present not only here but in many models of consumption under ambiguity aversion, robust control, and model uncertainty. It causes ambiguity aversion to typically lead agents to behave as though the world is persistent, and more than their point estimate implies (Hansen and Sargent (2010, 2015) and Bidder and Dew-Becker (2016); Fuster, Hebert, and Laibson (2011) obtain similar results in a slightly different setting). In past work, though, the information structure has been taken as exogenous. An important question is whether the results survive when agents are able to choose what to learn about. We find that they do not.

We show that the optimal learning policy is to choose information to align exactly with preferences. Since low-frequency fluctuations are most important for utility, the agent focuses attention on them. The effect is quantitatively large: the agent chooses signals that are 80 times more informative at the lowest than at the highest frequencies. That information then reduces their fears of model misspecification at low frequencies. In fact, the high precision at low frequencies perfectly cancels out the effect coming from them being most important for utility. At the optimum, agents

[3]The information acquisition that we study is a date-0 problem, and is thus different from dynamic learning, as in Abel, Eberly, and Panageas (2007, 2013), Wang (2009), Bansal and Shaliastovich (2010), Hansen and Sargent (2010), Ju and Miao (2012), and Collin-Dufresne, Johannes, and Lochstoer (2015)

neither over- nor under-extrapolate shocks to income. Our first theoretical result is thus that the past results on overextrapolation described above are sensitive to the specification of information.

It is not the case, though, that our agents make no mistakes. Since they do not have perfect knowledge of the income process, the models that the they use for decision making typically deviate from the true model driving income. The model's second prediction is that under the optimal information policy, agents tend to make relatively small mistakes about the low-frequency characteristics of the income process, whereas their models are much more likely to deviate from the truth at higher frequencies. There is a large literature that models information processing constraints as a source of deviations of behavior from full rationality. Our results suggest that such constraints should only lead to certain mistakes: if people get anything right, it will be the low-frequency characteristics of the income process.

The fact that agents have the least information at high frequencies also implies that is where there is most scope for disagreement. The model predicts that agents should tend to agree about the long-run effects of innovations to income, but may have different beliefs about the short-run dynamics.

As usual, if agents had complete information, consumption growth would be white noise. Any mistakes that the agents make in estimating the income process feed directly into deviations of consumption from a random walk. The finding that mistakes are larger at high than low frequencies then implies that deviations of consumption growth follow the same pattern. If, for example, income is predictable at high frequencies, consumption may inherit that predictability. Our results are therefore consistent with and can help explain empirical evidence on the apparent excess sensitivity of consumption to predictable changes in income.[4]

Other papers in the literature on rational inattention, including Reis (2006), Luo (2008), Maćkowiak and Wiederholt (2009), and Luo and Young (2010), also obtain excess sensitivity of consumption to income shocks, but through a different mechanism. In past rational attention models, agents obtain noisy signals about the state of the world.[5] That causes consumption to include errors that represent responses to noise in the signals and also to respond with a lag to true innovations in states. This paper, on the other hand, studies agents who obtain noisy signals about the model driving income. The excess sensitivity of consumption to predictable variation in income thus comes from errors in the agent's forecasting model, rather than errors in the estimate of the hidden state.

The fact that the mistakes that agents in our model make are mainly high- rather than low-frequency implies that they have relatively low utility costs (similar to Cochrane (1989), Luo (2008) and Kueng (2016)) and that the consumption process under the constrained-optimal information policy is close to that under full information. Specifically, the optimal information policy can be

[4] See Jappelli and Pistaferri (2010) and Kaplan and Violante (2014) for recent reviews.

[5] Luo and Young (2010) study a model with both rational inattention and a preference for robustness. The model uncertainty that the agents have in that model, though, is over the distribution of the shocks of the model, as in Hansen, Sargent, and Tallarini (1999) rather than directly over dynamics, as here, or in Bidder and Dew-Becker (2016) and Hansen and Sargent (2016).

shown to be close to one that minimizes the statistical distance between the agent's consumption process and the full-information optimum. Moreover, the welfare losses compared to the full-information benchmark from using the constrained optimal information policy are one to two orders of magnitude smaller than those obtained when an agent obtains equal information at all frequencies (a natural statistical benchmark, and similar to the setting studied by Fuster, Hebert, and Laibson (2011) and Bidder and Dew-Becker (2016)). The choice of information structures can thus have important effects on realized utility.

Finally, our results are also related to work on model approximation and heuristics.We show that the detail or complexity of the agent's final model is determined by the information that she acquires. The agent's model is simpler, in the sense that it is smoother and less variable, at frequencies where she gains less information and therefore relies more on her prior. Our work thus links directly to the literature on bounded rationality and heuristics in that it gives a description of how people might optimally construct simple approximations to complicated dynamic processes. Our paper differs from much past work on heuristics in that it focuses on how people might optimally learn about dynamic processes.[6] In that regard, it is closer to the literature on rational inattention in that people optimally choose what information to process, though the novelty here is the focus on uncertainty about dynamics instead of hidden states.

To summarize, then, this paper considers a consumption/saving problem in the face of nonparametric model uncertainty – the agent does not even know the lag order of the model she is supposed to estimate. We obtain a closed-form solution to the model, which yields insights into what aspects of the income process are most important for agents to learn about and what the implications are for observed consumption behavior. The analytic solution is, by itself, an important contribution of the paper, and should be useful in future analyses of optimal behavior under model uncertainty.

The next section describes the basic economic environment. We then proceed to describe and solve the attention allocation problem and finally examine its implications both analytically and through calibrations.

2 Environment and information

The problem we study consists of two stages. The second stage is a standard dynamic consumption problem in which an agent takes a certain model of her income process as given. In the first stage, the agent collects information and uses it to determine the model she uses to forecast income.

2.1 Economic environment

We study the problem of a person choosing how much to save and consume out of income. Consumers face an exogenous and untradeable stochastic income stream, Y_t, that they forecast using a

[6] See, e.g., Lipman (1995), Gigerenzer (2004), Payne and Bettman (2004), Gabaix (2016), and Schwartzstein (2014), among many others.

linear Gaussian model,

$$Y_t = \hat{a}(L) Y_{t-1} + \hat{b}_0 \hat{\varepsilon}_t \quad (1)$$

$$\hat{\varepsilon}_t \sim i.i.d.\ N(0,1) \quad (2)$$

where $\hat{a}(L)$ is a power series in the lag operator, L, with coefficients $\hat{a}_j$, and $\hat{b}_0$ determines the volatility of innovations to income. A particular model (1–2) will not in general be correct. The agents act as though $\hat{\varepsilon}_t$ is uncorrelated over time (i.e. treating it as a residual), but their assumption is only correct if $\hat{a}$ actually represents the true process driving dynamics (or at least the true projection coefficients).

Much of our analysis will apply to the Wold representation,

$$Y_t = \hat{b}(L) \hat{\varepsilon}_t \quad (3)$$

$$\text{where } \hat{b}(L) \equiv \frac{\hat{b}_0}{1 - L\hat{a}(L)}. \quad (4)$$

As with $\hat{a}$, we denote the coefficients in the power series $\hat{b}(L)$ using $\hat{b}_j$. From here on, then, rather than referring to models of income dynamics in terms of $\left\{\hat{a}(L), \hat{b}_0\right\}$, we will refer to them simply in terms of $\hat{b}(L)$. Since the distribution of $\hat{\varepsilon}_t$ is fixed, $\hat{b}(L)$ completely characterizes the statistical distribution of income. To be clear, though, the agent forecasts the future using only the past history of income. The $\hat{\varepsilon}_t$ are filtered from the income history (through $\hat{\varepsilon}_t = \hat{b}(L)^{-1} Y_t$, meaning that $\hat{b}$ is constrained to be invertible).

For simplicity, we assume that income is *actually* driven by a linear Gaussian process of the form (3) with coefficients b.

Assumption 1 *Income growth follows the process*

$$Y_t = b(L) \varepsilon_t \quad (5)$$

$$\varepsilon_t \sim i.i.d.\ N(0,1) \quad (6)$$

and b is unknown to agents.

The assumption that Y_t is a linear Gaussian process is not necessary for most of the results. The critical assumptions about the true income process are that it is second-order stationary and that it has a spectral density that is finite and bounded away from zero.[7] Our focus is on uncertainty about the dynamics of income, rather than about the distribution of its innovations. The latter question is obviously also interesting, but our goal here is to understand how consumption

[7] The spectrum can be thought of as the set of eigenvalues of the covariance matrix of an infinitely long history of the income stream. So the assumption that the spectrum is bounded from above and away from zero is equivalent to saying that those eigenvalues are positive and finite. If income has a unit root, which would imply an infinite spectrum, the analysis goes through identically applied to the first difference of income. We discuss that point further below.

responds to changes in income, and how well people understand the difference between permanent and transitory dynamics.

There is a single riskless saving technology with a constant gross return R.

Assumption 2 *Financial wealth, W_t, follows the process*

$$W_t = RW_{t-1} + Y_t - C_t \tag{7}$$

where C_t is consumption.

We also impose the usual transversality condition.

2.2 Model uncertainty and information structure

Knowledge of the income process requires estimation of the infinite collection of coefficients $\{b_j\}$. We study a nonparametric estimation process that is most natural to analyze in the frequency domain. The log spectral density of income, f, is defined as

$$f(\omega) \equiv \log \sum_{j=-\infty}^{\infty} \cos(\omega j) \operatorname{cov}(Y_t, Y_{t-j}). \tag{8}$$

There is a one-to-one mapping between the log spectral density and the autocovariances (since they are a Fourier transform pair), so the spectral density fully characterizes the income process. The behavior of f near $\omega = 0$ captures the low-frequency characteristics of income, while ω closer to π are higher frequencies. f is the true spectral density of income, while we denote alternative hypothetical spectra $\hat{f}$.

Given a particular spectral density $\hat{f}$, one may construct an associated Wold representation $\hat{b}$ (Priestley (1981) section 10.1.1). By focusing on the representation of the income process given by $\hat{f}$ rather than the one given by $\hat{b}$, we can more easily study how the agent learns about high- versus low-frequency dynamics of income. Since there is a one-to-one mapping between the two representations, no restrictions are imposed by working in the frequency domain.

2.2.1 Signals

The agents, before they begin to earn income, ask older people about their income histories. Their goal is to understand the dynamic properties of income – its variance and autocorrelations. Technically, we assume that the agents ask people about the coefficients from a regression of their income histories on cosines and sines that fluctuate at particular frequencies.[8] Intuitively, though, the goal is to understand whether income has large fluctuations at certain frequencies. If the agent is most

[8] If an agent wants to learn about fluctuations in income at some frequency ω, she will ask an older person to regress their income history, over a sample $t = 1, 2, ..., T$, on $\cos(\omega t)$ and $\sin(\omega t)$. The agent's observation of the log specturm from this person is then the log of the sum of the two squared coefficients.

interested in low-frequency dynamics, then they will ask more people about regression coefficients on low-frequency sines and cosines.

Mathematically, such questions correspond to asking people about parts of the log periodogram (the sample spectrum) of their income history. If all people have identical income dynamics, then the central limit theorem implies that the average of the answers that many people give yields a normally distributed signal about the log spectrum of income that has an error variance inversely proportional to the number of people they ask and uncorrelated across frequencies (see Brillinger (1981) section 5.2). The orthogonality is part of what allows us to solve the model and is the motivation for performing the analysis in the frequency domain. Estimates of ARMA coefficients are in general correlated – the Fourier transform is used in time series applications because it diagonalizes many problems.

For technical reasons, we assume that the agent gains information on the spectrum on the uniform discretization of $[0, \pi]$ given by $\omega_j = \pi j/n$ for $j \in \{1, ..., n\}$ (we will take n as large).

Assumption 3 *The agent receives signals* $\{x(\omega_j)\}_{j=1,...,n}$ *that are distributed as*

$$x(\omega_j) \sim N\left(f(\omega_j), \tau(\omega_j)^{-1}/d\omega\right) \tag{9}$$

where $d\omega \equiv \pi/n$ *and the errors are uncorrelated across frequencies.*

The function $\tau(\omega_j)$ is chosen by the agent and is proportional to the number of people that they ask about income fluctuations at frequency ω_j (we ignore discreteness in the number of people who are asked about the periodogram). The agent receives better information about the information that they ask more people about.

A natural benchmark is for the agent ask all people about their entire income history, rather than just a small number of regression coefficients. In that case, which is a natural statistical benchmark to which standard time series methods apply (e.g. Whittle (1951)), τ is constant across frequencies. We therefore take a constant τ as a natural benchmark that would appear if people did not allocate attention optimally.

Sims's (2003) model of rational inattention provides an alternative and equally important interpretation of the information structure. It is possible that complete information about the spectrum of income is available, but agents have trouble processing that information. Then the noise in the signals $x(\omega_j)$ represents cognitive errors that people make in interpreting the available information. The frequencies at which τ is larger are the ones the agent pays the most attention to. So we view the information structure in (9) as encompassing both models of costly information acquisition and also costly attention. This paper differs from the past literature on rational inattention in assuming that people are learning about the model driving income instead of state variables.

Under either interpretation, we assume that agents are limited in the total precision of their signals. That is, they face a constraint on the total number of observations of the periodogram (i.e. income regression coefficients) that they may ask about. Given the fact that $\tau(\omega_j)$ is proportional

to the number of people that the agent asks about fluctuations at frequency ω_j, we can say that they face a constraint on total precision (section 6.4 generalizes the constraint to allow for differential information costs across frequencies):

Assumption 4 *In choosing the precision of their signals, agents face the constraint*

$$\sum_{j=1}^{n} \tau(\omega_j)\, d\omega \leq \bar{\tau}. \tag{10}$$

2.2.2 Model plausibility

Since our goal is to understand how much information the agent desires to obtain, it is necessary for us to define her beliefs in the absence of information. We also aim, though, to minimize the informativeness of those prior beliefs. Given that the model space is infinite-dimensional, it is difficult to imagine that a person would have a fully defined prior. People likely cannot place a formal probability on every possible model, or even necessarily express a view about the relative likelihood of all possible pairs of models. We therefore specify prior beliefs as loosely as possible.

Our key assumption is that agents believe that the log spectrum is likely to be smooth in the sense that its differences across frequencies have limited variation. The smoothness prior is a belief in simplicity: agents believe that spectra typically are smooth across frequencies, rather than fluctuating wildly.

Following Shiller (1973), Akaike (1979), and Kitagawa and Gersch (1984, 1996), we model the belief in smoothness with a penalty on variability that is appended to the likelihood of the data.[9] Given assumption 3, the penalized log likelihood of the data given a model $\hat{f}$ is

$$PL\left(x \mid \hat{f}\right) = \underbrace{-\frac{1}{2}\sum_{j=1}^{n}\left(x(\omega_j) - \hat{f}(\omega_j)\right)^2 \tau(\omega_j)\, d\omega}_{\text{Data likelihood}} \underbrace{-\frac{\lambda}{2}\sum_{j=2}^{n}\left(\frac{\hat{f}(\omega_j) - \hat{f}(\omega_{j-1})}{d\omega}\right)^2 d\omega}_{\text{Roughness penalty}}. \tag{11}$$

$PL\left(x \mid \hat{f}\right)$ depends on two factors: the log likelihood for normally distributed data and a term encoding the belief in smoothness of the spectrum. Models are viewed as less plausible when they are rougher or more complicated in the sense of having larger average squared derivative. The most plausible models have perfectly flat spectra – white noise – while the least plausible have highly variable spectra.[10] That white noise is treated as the most plausible is also sensible from

[9] The smoothness prior is often explicitly jutified as a belief in simplicity. In Shiller (1973), which is the first application of such a prior, a justification is that "[i]n most applications...the researcher will feel that...the lag coefficients should trace out a 'smooth' or 'simple' curve." While Shiller's (1973) smoothness prior is stated in the time domain, those in Akaike (1979) and Kitagawa and Gersch (1985, 1989) are specified in the frequency domain in a manner almost identical to ours.

[10] An alternative way to penalize complexity in models would use the coefficients of the ARMA representation. We will see below, though, that the smoothness prior we impose here also ends up imposing smoothness on the AR and MA coefficients. The smoothness prior offers a general description of model complexity that allows the agent to use flexible specifications to model the income process, as opposed to imposing a perhaps more artificial constraint like

an information theoretic perspective since Gaussian white noise has the greatest Shannon entropy among all time series processes with a given variance.

The parameter λ controls the strength of smoothness prior. For any fixed λ, as the signal precision grows large, the smoothness penalty becomes irrelevant. Without the smoothness prior (when $\lambda = 0$), $\hat{f} = x$ is the maximum-likelihood estimate, which would imply that $\hat{f}$ has infinite variation and yield an inconsistent estimate of f, even as $n \to \infty$ (Wahba (1980)). Moreover, in that case, when τ shrinks and the agent's signals are less informative, the maximum-likelihood estimate of f becomes *more* variable, and hence more complex. We find it more plausible, instead, that when people have weak signals, they use simple models; complexity only arises in settings where people have a wealth of information.

The precision of the signals, τ, also controls the potential complexity of the model. When signals are more precise, so that τ is large relative to λ, the roughness penalty is relatively less important and the agent will consider more complex models. Conversely, when τ is small, the roughness penalty plays a larger role and agents will lean relatively more towards smooth and simple models.

In addition to the roughness penalty, we also assume that the agent is able to express a prior mean over possible models. In the absence of any information about the world, she believes that the average spectrum is flat at $\bar{f}$.

3 Preferences and the max-min-max problem

Studies of model uncertainty often take a min-max form in the sense that agents behave as though nature antagonistically chooses dynamics for income to minimize lifetime utility over consumption. In the min-max setup, the set of possible models from which the nature may choose is fixed exogenously. We extend the framework by adding a third layer in which the agent first chooses an information structure that constrains the set of models from which nature can choose.

The agent's choice of a consumption rule and nature's choice of an income process are modeled as the Nash equilibrium of a minimax game. We obtain optimal strategies by first finding the agent's optimal response to any model choice made by nature and then solving nature's problem of finding the model that minimizes the agent's optimized utility; that is, we apply the minimax theorem.

3.1 Consumption and utility conditional on an income model

Taking a particular dynamic model for income as given, people have time-separable utility with constant absolute risk aversion. Our analysis thus ignores wealth effects, but it is also more realistic than the assumption of quadratic utility over consumption (as in Hansen, Sargent, and Tallarini (1999), among others).

requiring agents to estimate only an AR(k) model for some exogenously limited k.

Definition 1 *Given a known income process, $\hat{f}$, agents choose consumption to maximize*

$$U\left(Y^t, W_{t-1} \mid \hat{f}\right) \equiv \max\left\{-\alpha^{-1}\log E_t\left[(1-\beta)\sum_{j=0}^{\infty}\beta^j \exp\left(-\alpha C_{t+j}\right) \mid \hat{f}\right]\right\} \tag{12}$$

conditional on the budget constraint (7) and where Y^t denotes the history of income up to period t.

The utility function over consumption is specified in its certainty equivalent form for the sake of simplicity in what follows.[11] That assumption is not necessary, though: we can also specify utility as being simply $E_t\left[\sum_{j=0}^{\infty}\beta^j \exp\left(-\alpha C_{t+j}\right) \mid \hat{f}\right]$ and obtain identical results.[12]

Lemma 1 *The unconditional expectation of utility is*

$$\begin{aligned} E\left[U\left(Y^t, W_{t-1} \mid \hat{f}\right) - (R-1)W_{t-1} \mid \hat{f}\right] &= -\frac{\alpha}{2}R^{-1}\left(1-R^{-1}\right)\hat{b}\left(R^{-1}\right)^2 \\ &\quad -\alpha^{-1}\log\frac{(1-\beta)}{1-R} - \alpha^{-1}\frac{\log\beta R}{R-1}. \end{aligned} \tag{13}$$

Lemma 1 characterizes the unconditional expectation of utility for a given income process $\hat{f}$ and initial level of wealth (the adjustment by $(R-1)W_{t-1}$ ensures that the expectation exists). The only term that differs across models is $\hat{b}\left(R^{-1}\right)^2$. $\hat{b}\left(R^{-1}\right)$ is the discounted sum of the impulse response function of Y_t to a shock to $\hat{\varepsilon}_t$. As usual, innovations to consumption growth under the optimal policy depend on innovations to the net present value (NPV) of income. $\hat{b}\left(R^{-1}\right)^2$ measures the variance of those innovations, and hence the variance of consumption growth. Utility is lower when the variance of consumption growth is higher.[13]

The information structure laid out in the previous section refers entirely to the log spectrum, but utility is derived in lemma 1 terms of the lag polynomial $\hat{b}$. We link the two through the following result.

Lemma 2 *For a log spectrum $\hat{f}$ that is bounded from above and below, where $\hat{b}(L)$ is the associated*

[11] See Van Nieuwerburgh and Veldkamp (2010) and Kacperczyk, Van Nieuwerburgh, and Veldkamp (2016) for a similar setup.

[12] Specifically, if we define $\tilde{U}\left(Y^t, W_{t-1} \mid \hat{f}\right) \equiv \max E_t\left[(1-\beta)\sum_{j=0}^{\infty}\beta^j \exp\left(-\alpha C_{t+j}\right) \mid \hat{f}\right]$, then expected utility is $\tilde{U}\left(W_0 \mid \hat{f}\right) = \exp\left(-\frac{\alpha}{2}R^{-1}\left(1-R^{-1}\right)\hat{b}\left(R^{-1}\right)^2 - \alpha^{-1}\log\frac{(1-\beta)}{1-R} - \alpha^{-1}\frac{\log\beta R}{R-1}\right)$, where we assume that the income history up to date 0 is equal to zero. Minimizing both $\tilde{U}$ and U is equivalent to maximizing $\hat{b}\left(R^{-1}\right)^2$, which is what will be relevant for decision making.

[13] Based on Lemma 1, our agents choose to learn about the variance of the shocks to the NPV of income. Our model of learning therefore may be relevant for more general utility functions like habit formation in which indirect utility is determined by the volatility of a net present value. Our analysis would be incomplete if people faced liquidity constraints that led other characteristics of the income process than its net present value to matter.

Wold (moving average) polynomial,

$$\log \hat{b}\left(R^{-1}\right)^2 = \frac{1}{\pi}\int_0^{\pi} Z(\kappa)\,\hat{f}(\kappa)\,d\kappa \tag{14}$$

$$\textit{where } Z(\kappa) \equiv 1 + 2\sum_{j=1}^{\infty}\cos(\kappa j)\,R^{-j}. \tag{15}$$

Lemma 2 gives us a powerful result: $\hat{b}\left(R^{-1}\right)^2$, the statistic that determines expected utility, is log-linear in the log spectrum. This result is the key innovation that will allow us to solve the model analytically.[14]

Lemma 2 shows that utility is always decreasing in $\hat{f}$, which is equivalent to saying that utility is decreasing in the variance of income growth. Moreover, though, utility depends on different frequencies differently, according to the function Z. The left-hand panel of figure 1 plots Z for an annual calibration with $R = 1.025$. $Z(\omega) > 0$ for all ω, it is bounded from above for $R > 1$, reaching its maximum at $\omega = 0$, and it is decreasing on $(0, \pi)$. When $R = 1$, Z is equivalent to the Dirac delta function. The mass of Z primarily lies on extremely low frequencies. So what matters for the agent's utility, through $\hat{b}\left(R^{-1}\right)^2$, is the magnitude of the spectral density at the very lowest frequencies. These characteristics of Z are robust features of the model, as they do not depend on any sort of detailed calibration – the only parameter affecting Z is the gross interest rate.[15]

3.2 Ambiguity aversion

Since we are interested in what information the agent desires to obtain, the key feature of preferences for our purposes is how the agent ranks different levels of uncertainty. One option is to simply assume that the agent acts as a pure Bayesian, treating uncertainty over shocks, ε, and models, f, symmetrically. That is, the future path of income depends on two sources of uncertainty, and the agent might simply integrate utility, $-\alpha^{-1}\exp(-\alpha C_t)$, over both of them.

A first problem with such analysis is that it would require specifying an infinite-dimensional prior that places a specific probability on every possible model. We have not specified such a prior here, and are doubtful that a person could actually articulate such beliefs. Furthermore, it

[14] Lemma 2 does not appear to have been previously noted in the literature, and we are not aware of any direct derivation from known results. It is a generalization of the Szegő–Kolmogorov formula for the innovation variance of a time series. Specifically, $\hat{b}(0)^2$ is the innovation variance, which the Szegő–Kolmogorov formula says is the geometric mean of the spectrum. The equation $\hat{b}(1)^2 = \exp\int \delta(\omega) f(\omega)\,d\omega$ for the Dirac delta function δ is also well known. So lemma 2 fills in $\hat{b}(x)^2$ for x between 0 and 1.

The innovation variance for the NPV of a time series arises naturally in many economic settings, such as the consumption/savings problem here, equilibrium macroeconomic models (Hansen and Sargent (1980, 1981)), models with generalized recursive preferences (Bidder and Dew-Becker (2016); Dew-Becker and Giglio (2016); Dew-Becker (2016)), the q theory of investment, and Calvo-type price setting. The appendix provides a proof.

[15] The analysis so far has assumed income is stationary. That assumption has no effects on our results. In the presence of a unit root, the analysis applies to the first difference of income. If $\hat{g}(L)$ is the Wold representation for the first difference of income, then $\hat{b}\left(R^{-1}\right) = \hat{g}\left(R^{-1}\right)/\left(1 - R^{-1}\right)$. The agent then can calculate $\log \hat{b}\left(R^{-1}\right)^2$ by using Lemma 2 applied to the log spectrum of income *growth* and subtracting $\log\left(1 - R^{-1}\right)$. The loading of utility on frequencies for the level of income is the same as for the first difference.

is well known that Bayesian analysis on infinite-dimensional parameter spaces (i.e. the space of possible spectra) is poorly behaved in the sense that maximum likelihood estimation is in general inconsistent and Bayesian confidence sets can have incorrect coverage.[16] It is also commonly argued that uncertainty about models is fundamentally different from uncertainty about the innovations ε, so it is not obvious that people would treat the two dimensions of uncertainty as being equivalent.[17]

We therefore model agents as ambiguity-averse over models. The agent chooses a consumption policy optimally under the assumption that nature simultaneously chooses a spectrum from the set of candidate models that minimizes lifetime utility subject to the constraint that it not be too implausible, as measured by the penalized likelihood (11).

The agent believes that the model chosen by nature will operate forever – we do not model dynamic learning.[18] The worst-case model is chosen timelessly in the sense that it minimizes the unconditional expectation of lifetime utility – it is chosen as the worst case from behind a veil of ignorance, not conditioning on any state variables. The agent, at the same time, chooses a consumption plan that is designed to be robust in an unfavorable world.

Since utility under the optimal consumption rule is a decreasing function of $\hat{b}\left(R^{-1}\right)^2$, we can say that nature simply maximizes $\log \hat{b}\left(R^{-1}\right)^2$ conditional on the penalized likelihood.

Definition 2 *Nature chooses f^w to minimize the unconditional expectation of the agent's lifetime utility, subject to a constraint on the penalized log likelihood of the data, $PL\left(x \mid \hat{f}\right)$:*

$$f^w(\cdot;\tau) \equiv \arg\min_{\hat{f}\in\mathcal{F}_n} \overbrace{-\sum_{j=1}^{n} Z(\omega_j)\,\hat{f}(\omega_j)\,d\omega}^{Expected\ utility} + \psi^{-1}\underbrace{\left(\frac{1}{2}\sum_{j=1}^{n}\left(x(\omega_j)-\hat{f}(\omega_j)\right)^2\tau(\omega_j)\,d\omega + \frac{\lambda}{2}\sum_{j=2}^{n}\left(\frac{\hat{f}(\omega_j)-\hat{f}(\omega_{j-1})}{d\omega}\right)^2 d\omega\right)}_{Penalized\ likelihood} \quad (16)$$

where ψ is a parameter determining how far nature can distort the model f^w from the data x and $\mathcal{F}_n$ denotes the set of left-continuous step functions on $[0,\pi]$ with respect to the discretization $\omega_1,...,\omega_n$.[19]

The notation $f^w(\cdot;\tau)$ emphasizes the fact that f^w depends on the precision, τ (we suppress the dependence on x for the sake of concision). f^w is the log spectrum that gives the agent the lowest expected lifetime utility subject to a constraint on plausibility. Intuitively, the ambiguity

[16] See Sims (1971), Chow and Grenander (1985), and Diaconis and Freedman (1986)

[17] E.g. Knight (1921), Ellsberg (1961), Hansen and Sargent (2007), and Machina and Siniscalchi (2014)

[18] We show that the agent focuses primarily on learning about low-frequency dynamics. Since the information that one has about very low frequency dynamics changes only slowly over time, learning is unlikely to be a major force in our setting.

[19] Technically, we say that nature minimizes the agent's expected utility from consumption conditional on the penalized likelihood not falling below some bound. ψ^{-1} is therefore a Lagrange multiplier.

aversion here says that agents make consumption choices that are meant to be robust against income dynamics that are set by nature to give them the lowest unconditional expectation for utility. The choice of the worst-case model is thus made in some sense timelessly: it does not depend on the recent history of income, but rather is chosen once and for all time to minimize utility integrated over all states of the world, ensuring time consistency. f^w is the solution to a linear quadratic optimization problem, so we can also expect in general that f^w exists and is unique.

Finally, as a technical matter, since the agent receives information at a discrete set of frequencies, we assume that the worst-case model is chosen by nature from the set of left-continuous step functions on the same discretization.[20]

The optimization problem in assumption 2 represents the minimization part of the max-min-max optimization we study. The final step is solving for optimal precision, τ.

3.3 Information choice objective

The agent's choice of signal precision, τ, must obviously be made prior to observing the signals, x. They thus have to maximize utility under the prior distribution of the spectrum, which, as noted above, is not completely defined. Note, though, that since f^w is the solution to a linear quadratic optimization problem, it is a linear function of x. That also then implies that $\log b^w \left(R^{-1}\right)^2$, which determines actual utility, is a linear function of x. So if the agents choose τ to minimize $\log b^w \left(R^{-1}\right)^2$, they only need a prior mean for f (and hence x), rather than its full distribution. To make it possible to solve the model, then, we assume that agents choose τ to minimize the expected value of $\log b^w \left(R^{-1}\right)^2$ under the prior distribution. The fact that this requires only a prior mean, rather than a full distribution, is consistent with our stated goal to minimize the specificity of the beliefs we assume the agent has.

That said, there is obviously a restriction being imposed here. Choosing τ to minimize some other function of $\log b^w \left(R^{-1}\right)^2$ (such as $\exp\left(b^w \left(R^{-1}\right)^2\right)$, which is what literally appears in expected utility) would require specifying a full prior for the spectrum (and hence x), would make the choice of τ depend on all the features of that prior, and does not yield a solvable model. The assumption that τ is chosen to minimize $\log b^w \left(R^{-1}\right)^2$ is thus the most parsimonious and requires the weakest assumptions about the agent's prior beliefs.

Definition 3 *The agent chooses the set of precisions,* $\{\tau(\omega_j)\}_{j\in\{1,2,\ldots,n\}}$*, to maximize expected*

[20] As $n \to \infty$, this assumption becomes minimally restrictive, and it does not rule out any economically reasonable specifications for the spectrum. It also allows us to replace the integral from lemma 2 with a finite sum. The large n assumption, though, does imply that people are able to obtain information about the spectrum at frequencies close to zero, which is equivalent to assuming that they have access to data from a very long time series. The economic implication of the restriction to the space of step functions is that agents do not fear that the spectrum of income explodes at frequency zero. The question of how people model behavior of the economy at frequencies at which they have essentially zero information (i.e. cycles lasting longer than we have modern economic data) is interesting but outside the scope of the present work.

utility subject to the information constraint

$$\{\tau^*(\omega_j)\} \equiv \arg\max_{\{\hat{\tau}(\omega_j)\}} E\left[-\sum_{j=1}^{n} Z(\omega_j) f^w(\omega_j;\hat{\tau})\, d\omega\right] - \theta \sum_{j=1}^{n} \hat{\tau}(\omega_j)\, d\omega \tag{17}$$

where the expectation is taken under the agent's prior and θ is the Lagrange multiplier on the information constraint from assumption 10.

4 Solution

The model is analytically solvable. We first characterize τ^*, then examine its implications for f^w.

4.1 Optimal information choice

Proposition 1 *The optimal information policy is*

$$\tau^*(\omega_j) = \underset{\textit{Cost of info.}}{\theta^{-1/2}} \times \underset{\textit{Ambiguity aversion}}{\psi^{1/2}} \times \underset{\textit{Utility weights}}{Z(\omega_j)} \tag{18}$$

Recall that the function Z measures how the level of the log spectrum, f, affects utility. Agents optimally gather information exactly in proportion to Z, learning the most about the frequencies that are most important for utility. In terms of the adversarial game with nature, the agent chooses precision to constrain nature most at the frequencies that are potentially most painful. Since τ also controls the potential complexity of f^w, the agent's choice of τ^* implies that models are most complex where Z is highest – very low frequencies.

While Z controls the shape of τ^*, θ and ψ determine its scale. An increase in the available precision $\overline{\tau}$ lowers θ, leading to more precision at all frequencies. ψ determines the extent to which nature is constrained by the penalized likelihood, i.e. how ambiguity-averse people are. Holding the shadow cost θ of precision constant, a decrease in ambiguity-aversion through ψ lowers the chosen precisions.

To see the implication of proposition 1 for noise in the signals at each frequency, the right-hand panel of figure 1 plots $Z(\omega)^{-1} \propto \tau^*(\omega)^{-1}$. The variance of the signals that the agents receive is a simple function of frequency, rising smoothly as the frequency increases.

4.2 The worst-case model

It is straightforward to solve for f^w from the first-order conditions for the nature's optimization in assumption 2. The solution can be obtained most easily by creating vectors (in boldface) of the form $\mathbf{f}^w(\boldsymbol{\tau}) \equiv [f^w(\omega_1;\tau), ..., f^w(\omega_n;\tau)]'$ (recall that the frequencies $\omega_j = \pi j/n$ are the uniform discretization of the interval $[0,\pi]$ on which the agent receives signals and that we think of n as

large). We define $diag(\cdot)$ to be an operator that creates a matrix with its argument on the main diagonal and zeros elsewhere.

Proposition 2 *For an arbitrary information policy τ, f^w is*

$$\mathbf{f}^w(\boldsymbol{\tau}) = \left(I_{n\times n} - \lambda diag\left(\boldsymbol{\tau}^{-1}\right) D\right)^{-1} \left(\psi diag\left(\boldsymbol{\tau}^{-1}\right) \mathbf{Z} + \mathbf{x}\right) \tag{19}$$

where $I_{n\times n}$ is an $n \times n$ identity matrix and D is a differencing matrix of the form

$$D \equiv \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & & \\ 0 & 1 & -2 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix} d\omega^{-2}. \tag{20}$$

We see that f^w is a linear function of x and Z. The worst-case spectrum is higher at frequencies where τ is smaller – there is more uncertainty about the spectrum – and where Z is larger – increases in the spectrum are more painful. Similarly, when ψ is larger, so that the agent is more ambiguity-averse, the worst-case model tilts more in the direction of Z.

The remainder of this section analyzes the properties of the worst-case model. We obtain the following results:

1. Without an optimal information (τ) policy, the worst-case model displays excessive persistence compared to the truth – people over-extrapolate shocks. But under optimal information (τ^*), that bias disappears.

2. Under the optimal policy, agents use models that tend to deviate from the truth more at high than at low frequencies.

3. Variation across agents in the worst-case model is inversely proportional to τ, implying that it is highest at high frequencies under the optimal policy.

4. The models agents use for forecasting are most complex where τ is high, hence at low frequencies.

4.2.1 The bias of f^w

Taking an expansion around an infinite level of precision, the appendix derives the following first-order approximation in the continuous limit of the problem ($d\omega \to 0$) for arbitrary τ:

$$E\left[f^w(\omega;\tau) - f \mid f\right] \approx \psi\tau(\omega)^{-1} Z(\omega) + \lambda\tau(\omega)^{-1} f''(\omega). \tag{21}$$

The expectation in equation (21) is conditional on the true data-generating model f instead of the agent's prior.

Equation (21) yields our first important result. In the statistical benchmark case where τ is constant across frequencies, we see that f^w is biased in the direction of $Z(\omega)$. Recall from figure 1 that Z is large at low frequencies and close to zero elsewhere. So under the statistical benchmark, the worst-case model has excessively high power at low frequencies, which means that it implies is more persistent than the truth (f). That result is almost exactly what is obtained in Bidder and Dew-Becker (2016), and is closely related to results in Hansen and Sargent (2010, 2016). Intuitively, since highly persistent models lead to the lowest utility, the agent naturally fears them.

Equation (21) also yields the second half of the first result, though, which is that under the optimal policy, τ^*, there is no systematic bias towards either under- or over-extrapolation. Specifically, we have under the optimal policy

$$E\left[f^w(\omega;\tau^*) - f \mid f\right] \approx \psi^{1/2}\theta^{1/2} + \lambda\tau^*(\omega)^{-1} f''(\omega). \tag{22}$$

Since $\tau^*(\omega) \propto Z(\omega)$, the frequencies that are most important for utility are also the ones that the agent learns the most about, thus constraining the worst-case model. The proportionality completely cancels Z out of the bias, leaving just a constant. When f^w deviates from f by only a constant, that means that they have identical autocorrelations and differ only in the conditional variances. For example (ignoring the effects of f'' for the moment; i.e. for small λ) if income follows an AR(1) process with persistence ρ, then $E[f^w]$ is the log spectrum for an AR(1) also with persistence ρ, but with innovations that have a greater variance.

Equation (22) is a key result of the paper. It shows that endogenous learning can completely eliminate overextrapolation. Intuitively, ambiguity averse agents tend to focus on models with excessive persistence because they are associated with low utility. But that fact also causes them to obtain the most information about those frequencies, thus canceling out the effect of ambiguity. That said, this result is obviously relevant only in settings where there is information to acquire. When learning about individual income dynamics, where one can look to the income paths of friends of relatives, there is likely to be substantial information available. When looking at the dynamics of the aggregate economy, though, we might expect there to be less information available, making learning less important and making equation (21) more relevant. On the other hand, under the rational inattention interpretation of the model, we would expect this result to apply in all settings (since in that case the binding constraint is not in the available information, but in the ability of people to process it).

So far we have ignored the term $f''(\omega)$. That part of the formula is driven by the agent's smoothness prior and represents our second result, that mistakes tend to be higher at high frequencies. Intuitively, in the face of noisy data, agents estimate the spectrum of income by smoothing information across frequencies. At frequencies where there is less information, or when the prior belief in smoothness is stronger, the range of frequencies that are smoothed across gets wider. Since

$f^w(\omega;\tau)$ is a convex combination of the data x local to ω, it is biased upward when $f'' > 0$ and downward when $f'' < 0$. Intuitively, if there is a narrow peak in f, a simple model will tend to smooth the peak out, and thus be biased downward.

So equation (22) predicts that people make relatively larger errors at high frequencies, where they have weaker information. At high frequencies, they smooth out the spectrum relatively more, essentially ignoring its details. At low frequencies, though, agents have high precision, making $\lambda\tau^*(\omega)^{-1} f''(\omega)$ small, and implying that they are better able to capture the details of the spectrum there.

4.2.2 Local approximation

Equation (19) shows that f^w is linear in x and Z, but it is difficult to interpret the exact weights in that function. The appendix therefore develops a simple approximation that reveals more detail about how the worst-case model is constructed. We derive a first-order approximation for $f^w(\omega_j;\tau)$ in terms of τ^{-1} around zero (i.e. for infinite precision) and in the continuous limit where $d\omega \to 0$, that we denote f^w_{CL}.

The appendix shows that the f^w_{CL} takes the form

$$f^w_{CL}(\omega;\tau(\omega)) \equiv \int_{-\infty}^{\infty} K(\omega - m;\tau(\omega),\lambda)\left(\psi\tau(\omega)^{-1} Z(\omega - m) + x(\omega - m)\right) dm \tag{23}$$

$$\text{where } K(\omega - m;\tau(\omega),\lambda) \equiv \frac{1}{2}\tau(\omega)^{1/2}\lambda^{-1/2}\exp\left(-|m|\,\tau(\omega)^{1/2}\lambda^{-1/2}\right). \tag{24}$$

where the integration from $-\infty$ to ∞ is performed by reflecting x at intervals of 2π to create an extension to the whole real line.

Equation (23) shows that the approximate solution, $f^w_{CL}(\omega;\tau(\omega))$, is a kernel smoother, K, applied to a mixture of the observed signals, $x(\omega)$, and, as above, the potential bias towards excessive extrapolation, $\psi\tau^{-1}Z$. As before, when $\tau^{-1}Z \propto 1$, that bias becomes constant across frequencies, having no effect on autocorrelations.

The kernel, K is simply an exponential function reflected across the origin, with the property that $\int_{-\infty}^{\infty} K(\omega;\tau,\lambda)\,d\omega = 1$. (23) thus implies that f^w_{CL} depends on the data, x, but because of the smoothness prior, $f^w_{CL}(\omega;\tau) \neq x(\omega)$. Instead, it optimally smooths across values of x. The kernel is wider, in the sense that its mass declines more slowly for frequencies away from ω, when λ is larger – the agent has a stronger prior on smoothness – or τ is smaller – each data point x is less informative (in the limit when $\tau \to \infty$, the kernel becomes equivalent to the Dirac delta function). The presence of the kernel comes from the term involving f'' in the expression for f^w.

For the next two results, we describe the behavior of f^w by analyzing the behavior of the limiting first-order approximation f^w_{CL}. We then use numerical simulations to confirm that those approximate limiting results are reasonable descriptions of the actual behavior of the model in practice.

4.2.3 Variance of f^w

Using standard results on kernel smoothers, we have

$$Var\left[f_{CL}^{w}\left(\omega;\tau\left(\omega\right)\right)\right] \approx \frac{1}{4}\tau\left(\omega\right)^{-1/2}\lambda^{-1/2}\left(1+1\left\{\omega \bmod \pi = 0\right\}\right) \tag{25}$$

where $1\{\cdot\}$ is the indicator function. As above, this result follows from a first-order approximation in terms of τ^{-1} around zero. When τ is higher, the variance of x at each frequency is smaller. If the kernel used to construct $f_{CL}^{w}\left(\omega;\tau\left(\omega\right)\right)$ stayed fixed as $\tau\left(\omega\right)$ varied, then the variance of $f_{CL}^{w}\left(\omega;\tau\left(\omega\right)\right)$ would be proportional to $\tau\left(\omega\right)^{-1}$. But the kernel also changes, becoming narrower when $\tau\left(\omega\right)$ is higher. Intuitively, when each data point is more informative, the agent's estimate uses fewer points. So the variance of $f_{CL}^{w}\left(\omega;\tau\left(\omega\right)\right)$ ends up being proportional to $\tau\left(\omega\right)^{-1/2}$, rather than $\tau\left(\omega\right)^{-1}$. That is our third result from above: the variance across people of estimates of the spectrum is high where τ is low – high frequencies.

The term $\left(1+1\left\{\omega \bmod \pi = 0\right\}\right)$ shows that the variance of the estimate doubles at the boundaries. For ω close to zero, $f^{w}\left(\omega\right)$ is mostly estimated based on information to the right of ω, whereas for interior frequencies, $f^{w}\left(\omega\right)$ has effectively twice as much information to use since it can smooth data from frequencies on both sides of ω.

The parameter λ determines how strongly the agent holds the prior belief that the log spectrum is smooth. When λ is large, the desire for smoothness causes f_{CL}^{w} to be estimated using a wider kernel. λ therefore induces a bias/variance tradeoff. Higher λ, or greater model simplicity, reduces variance, since $f_{CL}^{w}\left(\omega_j;\tau\left(\omega_j\right)\right)$ is estimated using information from a wider range of frequencies, but it also increases bias.

We can think of bias as representing systematic mistakes that people make and variance as representing disagreement. That is, even if the agents all receive independent draws of x, they all have the same bias. So the model predicts that agents tend to make correlated mistakes, in the sense of using incorrect models, at places where the spectrum is most curved. Those mistakes will be small, though, where τ is large: the lowest frequencies.

Similarly, the scope for disagreement depends on how noisy the signals are. In regions where the signals have very little noise, i.e. low frequencies, agents will all receive very similar signals, and thus tend to agree about the spectrum. On the other hand, the signals are noisier at high frequencies, so the model predicts that agents should disagree most at those frequencies. The model thus predicts that bias and variance are positively correlated, both peaking at high frequencies.

4.3 Attention and model complexity

A simple way to measure the complexity of the worst-case model is to examine how strongly correlated it is across frequencies. When the worst-case spectrum is more strongly correlated, then it is relatively smoother across frequencies (equivalently, the entropy of the joint distribution of f^w across frequencies is smaller). The appendix shows that the correlation across frequencies can be

approximated as

$$Corr\left(f_{CL}^{w}\left(\omega;\tau\left(\omega\right)\right),f_{CL}^{w}\left(\omega+\Delta\omega;\tau\left(\omega\right)\right)\right)\approx1-\left|\Delta\omega\right|^{2}\tau\left(\omega\right)\lambda^{-1}. \tag{26}$$

The approximation is valid for large τ and small $\Delta\omega$. As we would expect, the correlation between points is decreasing in their distance. The rate at which the correlation declines depends on precision and the smoothness prior. When people have a stronger belief in smoothness – λ is higher – f^w is less variable across frequencies and thus smoother. On the other hand, when people have more information at each frequency, the correlation across frequencies declines more rapidly.

So at the frequencies where the agent obtains more precise estimates, i.e. low frequencies, the estimated spectrum, f^w, tends to vary more across frequencies. In other words, the spectrum is more detailed – in the sense that it is allowed to vary more in response to the data – at the frequencies where the agent gathers the most information. Equation (26) further develops the intuition we have suggested previously about the relationship between τ and model complexity: optimization over information is optimization over complexity.

This result is the fourth key insight of our analysis. There is a tight link between model complexity and information acquisition. The smoothness prior of the agent we are studying can be thought of as a prior on model simplicity: the agent believes that spectra typically vary smoothly, making them correlated across frequencies. So when the agent has weak data, with low τ, she relies primarily on the prior and uses a smooth and simple model. But when the data is sufficiently informative, the likelihood eventually wins out over the prior and the agent can consider models of high complexity. And that relationship is tight: it is precisely τ that determines how the complexity of the model varies across frequencies.

The optimal information policy then seems natural: agents devote their attention, and thus the complexity of their models, to the frequencies that matter most for utility.

5 The behavior of consumption

This section characterizes consumption dynamics in our framework and discusses the relationship with empirical work. The main result is that under the optimal information policy, consumption tracks income excessively closely at high frequencies, but in the long-run the consumption policy is close to the full-information optimum. A suboptimal information policy that acquires equal information at all frequencies predicts the opposite: consumption does not track transitory fluctuations in income, but it has long-term predictability. Those results are derived analytically here and in simulations in the next section.

5.1 Consumption

Suppose the agent chooses consumption as though income is driven by the model b^w, where the true income process is b. The appendix shows that consumption growth then follows the process

$$\Delta C_t = \left(1-R^{-1}\right) b^w\left(R^{-1}\right)\varepsilon^w_{t+1} + \frac{\alpha}{2}\left(1-R^{-1}\right)^2 b^w\left(R^{-1}\right)^2 + \alpha^{-1}\log\beta R \quad (27)$$

$$\text{where } \varepsilon^w_{t+1} \equiv \left(b^w_0\right)^{-1}\left(Y_t - a^w(L)\,Y_{t-1}\right) = b^w(L)^{-1}\,Y_t \quad (28)$$

where b is the true model and $\Delta \equiv 1-L$ is the first-difference operator. This is a general result that simply assumes that consumption is chosen optimally as though b^w drives the income process. It requires no assumptions about how b^w is chosen.

The second term in equation (27) is due to precautionary saving: when the riskiness of the economy is higher, consumption grows more quickly. The first term is more interesting, though, determining dynamics. In the case where $b^w = b$, the filtered shocks, ε^w are equal to the true shocks, ε, and consumption follows a random walk with innovations equal to the innovation in the annuity value of the NPV of future income, $\left(1-R^{-1}\right) b\left(R^{-1}\right)\varepsilon_{t+1}$.[21]

When the agent uses a model that differs from the truth, though, ε^w_{t+1} is no longer an i.i.d. process and consumption growth is no longer uncorrelated over time. That is, unless $b^w(L) \propto b(L)$, the estimated shocks, ε^w, are serially correlated, which leads to (suboptimal) serial correlation in consumption growth.

To better understand the implications of this result for the behavior of consumption growth, we can write the log spectrum of consumption growth as

$$f^w_{\Delta C}(\omega) = \log\left(\left(1-R^{-1}\right)^2 b^w\left(R^{-1}\right)^2\right) + f(\omega) - f^w(\omega). \quad (29)$$

When the agent knows the true model, $f^w_{\Delta C}$ is perfectly flat – consumption is a random walk. But in general the agent does not know the true model. For example, if the true spectral density has a peak at some frequency but the worst-case spectrum does not, then $f^w_{\Delta C}$ will inherit the same peak through the term $f(\omega) - f^w(\omega)$. That is, features of the income spectrum that the agent "ignores" in the sense that they do not appear in f^w are passed through to the spectrum of consumption growth.

As shown by equation (22), the difference between f^w and f systematically deviates from zero more at higher frequencies because τ^* is lower there. Specifically,

$$E\left[f^w_{\Delta C}(\omega) \mid f\right] \approx \log\left(\left(1-R^{-1}\right)^2 b^w\left(R^{-1}\right)^2\right) - \psi^{1/2}\theta^{1/2} - \lambda\tau^*(\omega)^{-1} f''(\omega) \quad (30)$$

Deviations on average of consumption growth from white noise are caused by the term $\lambda\tau^*(\omega)^{-1} f''(\omega)$ – they occur where the true spectrum is most complex (in terms of curvature) and where the agent

[21] More generally, consumption is a random walk when $b^w \propto \bar{b}$. If the factor of proportionality is not equal to one, b^w encodes the true dynamics, but with an incorrect volatility. The agent then has a suboptimal mean growth rate of consumption, but the response of consumption to shocks is the same as under the optimal plan.

acquires the least information.

Since τ^* is smallest at high frequencies, all else equal, that is where we are most likely to see deviations of consumption growth from white noise. For example, agents might rationally ignore variation in income over the course of a year (e.g. tax refunds or hitting the social security tax cap; see below). The model predicts then that consumption would fluctuate in sync with those shifts in income, since people would take them as surprise increases.

Conversely, it is less likely for the agent to make mistakes at low frequencies, since that is where τ^* is largest, so $f^w_{\Delta C}$ is predicted be more flat for small ω. This flatness implies that consumption growth over long, disjoint periods should be close to uncorrelated.

5.2 Empirical evidence

Since the optimal information policy implies that people learn the most about low-frequency features of the income process, it says that deviations of consumption growth from white noise should be observed primarily at high frequencies. Specifically, if the agent's model of income dynamics, $f^w(\omega;\tau^*)$, is flat at high frequencies, then any variation in the shape of the true spectrum passes directly into consumption. The shape of the spectrum of $f^w_{\Delta C}(\omega;\tau^*)$ will typically be similar to that of $f(\omega)$ at high frequencies as the model predicts that people use simple (flat) models there.

Another way to build intuition for that prediction of the model is to note that high-frequency shocks also have relatively small effects on the net present value of income compared to more persistent shocks (which is why the function Z is relatively small at high frequencies). So the model essentially predicts that people spend excessively out of relatively small high-frequency increases in income compared to the larger low-frequency shocks.

Those predictions of the model are consistent with recent empirical evidence. Parker (1999) and Souleles (1999) provide classic evidence on the response of consumption to predictable changes in income due to the tax code (the cap on social security taxes and tax refunds, respectively). Kaplan and Violante (2014; see references therein) review extensive evidence on the effectiveness of fiscal stimulus payments, finding that people tend to spend approximately 25 percent of these transitory payments in the quarter that they are received, even though the standard frictionless model would imply that they should spend a fraction near the level of the real interest rate (i.e. 1 percent or less per quarter). Moreover, these responses occur even among people with high incomes, who are less likely to be liquidity constrained (see also Kueng (2016)).

Kaplan and Violante explain the empirical evidence by arguing that when people hold illiquid assets, their consumption is excessively sensitive to transitory shocks because the benefit of smoothing is smaller than the cost of adjusting the stock of illiquid assets (e.g. housing). The intuition behind our results is similar to theirs (and also that of Cochrane (1989)) in that our results are also driven by the relatively small welfare benefit of smoothing transitory shocks. We differ in emphasizing the cost of learning about high-frequency dynamics, as opposed to assuming that saving is costly. Kaplan and Violante (2016) note that their model is consistent with the finding

of Hsieh (2003) that consumption seems to respond relatively more to small than to large income shocks. That intuition is consistent with our argument that it is most natural for people to learn about shocks that have large effects on human wealth.

While the key source of variation for Kaplan and Violante (2014) is the size of shocks to income, for us it is their duration. Consumption should be most sensitive to high-frequency variation in income, while at longer horizons, it should be closer to white noise. The empirical research finding violations of the permanent income hypothesis typically finds that evidence at high frequencies.

Cochrane and Sbordone (1988) examine the joint relationship between aggregate consumption and output at long horizons and find that consumption helps forecast future output growth, but output does not help forecast consumption (nor do lags of consumption itself), implying that consumption growth is approximately white noise at long horizons. In other words, our model is consistent with the view that consumption growth may deviate from white noise and respond excessively to income in the short-term, but at longer horizons it is well described as white noise.

An alternative way to test the model would be to directly ask people what they are willing to pay for information. If they are at the optimum τ^*, then information is equally valuable at all frequencies. On the other hand, under the standard models of ambiguity aversion without endogenous information acquisition, people would value low-frequency information most highly and be willing to pay the most for it.

5.3 Relationship with the full-information optimal consumption rule

Our information-constrained agent uses a consumption rule that is suboptimal to the extent that $b^w(L)$ differs from $b(L)$. b^w is not chosen to directly generate a path for consumption that necessarily maximizes realized utility. We now show, though, that the agent's worst-case optimization problem is closely related to an optimization that approximates the correct consumption rule.

Remark 1 *A second-order expansion of the Kullback–Leibler (KL) divergence between the full-information rational expectations consumption process and that used by an agent with model f^w around the point $f^w = f$ is*

$$KL\left(f_{\Delta C}; f^w_{\Delta C}\right) \approx \frac{1}{4\pi}\int_0^{2\pi}\left(\left(Z(\omega)-1\right)^2 + 2\left(R^{-1}\frac{\alpha}{2}\left(1-R^{-1}\right)\right)^2 Z(\omega)^2\right)\left(f^w(\omega)-f(\omega)\right)^2 d\omega. \tag{31}$$

The KL divergence is a likelihood-based measure of the deviation between the two random processes (one interpretation is that it measures how likely one would be to reject the hypothesis that consumption is driven by one process after observing data generated by the other). We see that the KL divergence weights squared errors in the model f^w by a quadratic function of $Z(\omega)$. As long as R is close enough to 1, this weighting function is strictly maximized at $\omega = 0$, meaning that reducing the distance between f^w and f at low frequencies reduces the KL divergence the most. The optimization problem that our agent solves involves minimizing squared errors in $f^w(\omega)$

weighted by $\tau^*(\omega)$, and theorem 1 shows that $\tau^*(\omega) \propto Z(\omega)$. The estimations of the agent and of someone minimizing KL divergence both involve using the weights given by Z to put more emphasis on the precision of the estimate at low frequencies.

6 Numerical example

To help illustrate the key mechanisms in the model, we now consider a numerical example. The left-hand panel of figure 2 plots the log spectrum of the data-generating process for income. It is specified as a combination of an AR(2), which generates a peak at middle frequencies, and an ARMA(1,1) that is calibrated to generate a peak of the same size at frequency zero. The calibration thus allows us to see how an agent models deviations of income from white noise at different frequencies.

We set $n = 4000$ and $\beta = 0.975$. The parameters θ, λ, and ψ are chosen in order to ensure that the agents make non-trivial mistakes in modeling consumption and that the behavior is visibly different across the two policies for τ described below.[22] The calibration is meant to illustrate the qualitative behavior of the model rather than match specific quantitative data.

To understand the effects of optimal information acquisition, we examine two specifications for τ: the first is the optimum derived above, τ^*, which is proportional to $Z(\omega)$; the second specification is the statistical benchmark that sets $\tau(\omega)$ to be constant at the mean of τ^*:

$$\tau^F(\omega_j) = \tau^F \equiv n^{-1} \sum_{i=1}^{n} \tau^*(\omega_j). \tag{32}$$

The choice of the mean for τ^F implies that it has the exact same information cost as τ^* (from (17)). Note, though, that since precision is the inverse of variance, the average variance of the errors across frequencies is in fact much smaller under τ^F than under τ^*. In our benchmark calibration, the average variance under τ^*, $n^{-1}\sum_{j=1}^{n} \tau^*(\omega_j)^{-1}$, is equal to 971, but the average variance under τ^F is only 10.0. In exchange, though, the variance of the errors under τ^* is far smaller at the lowest frequencies: at $\omega_1 = \pi/n$, $\tau^*(\omega_1)^{-1}$ is, at 0.049, 200 times smaller than $\tau^F(\omega_1)^{-1}$.

The majority of our analysis focuses on the average worst-case models:

$$\bar{f}_*^w(\omega) \equiv E\left[f^w(\omega;\tau^*) \mid f\right] \tag{33}$$

$$\bar{f}_F^w(\omega) \equiv E\left[f^w\left(\omega;\tau^F\right) \mid f\right]. \tag{34}$$

Figure 2 plots $\bar{f}_*^w$ and $\bar{f}_F^w$. The two log spectra are rather different. $\bar{f}_*^w$ matches f almost perfectly at the very lowest frequencies, but it does a poor job of matching the middle-frequency peak in $\bar{f}$ and also deviates substantially at higher frequencies. $\bar{f}_F^w$ has the opposite behavior: it matches the middle-frequency peak and high-frequency behavior well, and in fact matches f well at

[22] $\psi = 1/12000$; $\lambda = 0.00075$; $\theta = 2520$.

almost all frequencies, but does relatively worse at low frequencies. Overall, $\bar{f}_F^w$ has a much better fit than $\bar{f}_*^w$, with a root mean squared error that is 79 percent smaller.

The right-hand panel of figure 2 plots the lag polynomials, b, $\bar{b}_*^w$, and $\bar{b}_F^w$, associated with the log spectra f, $\bar{f}_*^w$, and $\bar{f}_F^w$, respectively (note that the lag polynomials are equivalent to impulse-response functions of income to the filtered innovations ε, $\bar{\varepsilon}_*^w$, and $\bar{\varepsilon}_F^w$). $\bar{b}_*^w$ clearly does not match the oscillatory dynamics in the income process, while the lag polynomial for the suboptimal information policy, $\bar{b}_F^w$, does. That behavior is driven by the fact that $\bar{b}_*^w$ does a poorer job of matching the middle-frequency peak in the spectrum of income, which is what causes the oscillations in income. The figure shows, then, that the greater smoothness of $\bar{f}_*^w$ also translates into smoothness in the associated lag polynomial.

6.1 Consumption

Figure 2 shows overall that estimation using information that is equally precise across all frequencies does a better job of matching the dynamics of income in a mean-squared-error sense. More relevant for the agent in the model, though, is what the different estimation methods imply for the behavior of consumption growth and asset returns (recall from above that they have identical dynamics). Figure 3 therefore plots the log spectra of consumption growth under the various models. The budget constraint forces the spectra to all be approximately equal at very low frequencies. At higher frequencies, though, we see that $\bar{f}_*^w$ actually seems to fit better than $\bar{f}_F^w$. The root mean squared error for the worst-case consumption rule compared to the full-information rule is 0.32, compared to 0.41 for the rule generated under τ^F.

Since $\bar{f}_F^w$ is much smaller than f at the very lowest frequencies, $\bar{f}_{F,\Delta C}^w$ has a large peak at very low frequencies, and the spectrum everywhere else is shifted downward. Intuitively, agents using the flat information policy, τ^F, underreact to income shocks because they underestimate their persistence. The short-term underreaction is then followed by an increase in wealth and consumption in the long-run, inducing a large persistent component in consumption growth. In the long-run, the consumption response must be the same under all models due to the budget constraint (i.e. the response of the NPV of consumption to a shock must be the same as the response of the NPV of income for the budget constraint to hold). The agents using the optimal information policy, τ^*, do a much better job of estimating low frequency dynamics, so the spectrum of consumption growth, $\bar{f}_{*,\Delta C}^w$, has a much smaller peak at frequency zero.

To see how the low-frequency errors affect the behavior of consumption growth in the time domain, the right-hand panel of figure 3 plots the impulse response of the level of consumption to a unit shock to ε_t (i.e. a true innovation, not a filtered one) under the three consumption rules along with the cumulative impulse response of income (multiplied by $\left(1 - R^{-1}\right)$). As we would expect, the response of consumption under the full-information rule is flat: the permanent income hypothesis holds. We see that the response of consumption is approximately equal to the cumulative increase in income. The line for consumption under the optimal information policy shows that

it inherits much of the high-frequency behavior of income, rising and falling exactly as income does. It does not include the persistent component in income, though – consumption immediately jumps to approximately its long-run level, but the fluctuates around that level excessively. So the consumption policy is "right" in the long-run, but it is excessively sensitive to transitory variation in income in the short-run.

The behavior of a person using the model $\bar{f}_*^w$ is again notably different from one using $\bar{f}_F^w$. The latter model does a better job of eliminating high-frequency fluctuations in consumption, but at the cost of inheriting the low-frequency behavior of income. The initial response of consumption under $\bar{f}_F^w$ is too small, and consumption slowly drifts upward over the 40 periods of the IRF plotted here. So the τ^F policy, counter to what we observe empirically, eliminates the sensitivity of consumption to transitory fluctuations in income, but causes consumption growth to deviate from white noise at long horizons. This result argues that empirically, τ^* is a better description of consumption behavior than a setting where agents do not choose information optimally, τ^F.

What this section ultimately shows is that the optimal information policy, while it is much less accurate than τ^F at most frequencies, does a better job of matching the optimal consumption policy (and the data) because it fits the low-frequency behavior of income better, exactly as suggested by the result on the KL divergence in remark 1. It also confirms the analytic results above that the optimal information policy does a good job of generating consumption growth that is close to white noise in the long-run, but that it causes consumption to be excessively sensitive to variation in income in the short-run.

6.2 Disagreement

The results so far have focused on the behavior for the average value of $\bar{f}_*^w$. But one interpretation of the model is that different agents receive different signals about the dynamics of the economy. To see the what the model implies for disagreement, the left-hand panel of figure 6 plots the cross-sectional standard deviation of f_*^w and f_F^w. Not surprisingly, given the shapes of τ^* and τ^F, disagreement is approximately constant across frequencies under τ^F, while it increases strongly with frequency under τ^*. There are also endpoint effects: because agents only have information on the bounded interval $[0, \pi]$, there is effectively less information local to the endpoints than there is on the interior of the interval. That effect is clearest under τ^F, where the cross-sectional standard deviation is 41 percent higher at the endpoints than the interior of the interval (note that this almost exactly consistent with equation (25), which says that the standard deviation should be $\sqrt{2}$ times higher at endpoints than in the interior).

The right-hand panel of figure 6 plots the cross-sectional standard deviation of the IRF of the level of consumption following a unit standard deviation shock to ε_t. The standard deviation is reported as a fraction of the final response of the level of consumption. In both specifications, this measure of disagreement declines with the horizon, so that the standard deviation across individuals in their initial consumption responses to the shock is much larger than the standard deviation of

their long-run responses. The initial cross-sectional standard deviation is two times larger under τ^* than under τ^F.

6.3 Realized utility

The agents that we study have ambiguity-averse preferences, and the behavior analyzed above under the τ^* information structure is optimal for them under the model. A natural question, though, is how their *realized* utility – i.e. the utility over the goods that they actually consume – behaves under the various forecasts for income. Since the model is fully linear and Gaussian, it is straightforward to calculate the mean of discounted realized utility,

$$E\left[-\alpha^{-1}\sum_{j=0}^{\infty}\beta^j \exp\left(-\alpha C_{t+j}\right) \mid f, W_{t-1}\right] \tag{35}$$

where the mean here is taken conditional on the true income process. Table 1 lists the loss in mean realized utility under τ^* and τ^F compared to the optimal consumption policy measured in terms of an equivalent percentage shift in lifetime consumption. We set initial wealth to support consumption of 1, so that absolute risk aversion corresponds to relative risk aversion in the first period.

Risk aversion	τ^*	τ^F	$\tau^{b\left(R^{-1}\right)}$
4	0.0019	0.0319	0.0562
6	0.0032	0.0824	0.1020
10	0.0060	0.3412	0.2469
15	0.0142	1.4963	0.7285

Table 1. Utility losses compared to optimal policy (percent of lifetime consumption)

Table 1 shows that, at least in relative terms, τ^* delivers enormous improvements in realized utility over τ^F: the utility losses relative to the optimal policy are up to two orders of magnitude smaller. Intuitively, the reason for this is that the agent using τ^F sets the mean consumption growth rate to the wrong level and gets low frequency dynamics wrong, inducing large and persistent errors in the level of consumption.

Mean consumption growth depends on the size of the precautionary saving effect, which itself depends on $\hat{b}\left(R^{-1}\right)$. Because an agent using τ^F tends to use a value of $\hat{b}\left(R^{-1}\right)$ further from the truth than an agent using τ^*, the mean consumption growth rate tends to be farther from optimal. Moreover, though, since τ^F yields a poor estimate of income dynamics at low frequencies, the errors in the associated consumption policy are long-lasting and large, whereas under τ^* consumption tends to fluctuate at high frequencies close to the optimal path.

Since the NPV of consumption is the same under all the policies, the effects on utility are all second order – they come only from rearranging consumption over time, and even with risk aversion

of 15, the loss in utility using τ^F is equivalent to only 1.5 percent of lifetime consumption.[23] While that effect is not enormous, it is still over 100 times larger than the loss from using τ^*. Kueng (2016) provides evidence consistent with our model that people consistently make mistakes in consumption, but that they tend to be smaller when they have larger utility costs.

The fourth column adds results from a model we have not discussed so far. As noted above, if the goal is simply to constrain nature's choice of the worst-case model, the key moment for an agent to learn is $b\left(R^{-1}\right)$. If the agent knows only that moment and nothing else about the behavior of income, then the model with the highest likelihood under the smoothness prior would have a flat spectrum at the level $b\left(R^{-1}\right)^2$. The advantage of such a model is that it yields mean consumption growth identical to the optimum. It gets consumption dynamics completely wrong, though, treating income as though it is uncorrelated over time.

The column headed "$\tau^{b\left(R^{-1}\right)}$" reports realized utility for a person who uses such an income model. We see that utility is similar to what is obtained under τ^F. That result shows that τ^* provides superior realized utility not just because it yields the correct mean consumption growth rate but also because it yields a good estimate of income dynamics (at least at low frequencies).

So table 1 shows two things. First, the losses in realized utility from a suboptimal consumption plan are generally small – in our case less than 2 percent of lifetime utility. But second, the losses can be cut to an enormous degree not even by gathering more information, but rather just allocating attention to the right pieces of information.

6.4 Extension: frequency-dependent information costs

In the baseline model, assumption 4 implies that agents have equal ability to learn about all frequencies. That assumption is most natural in the limited attention interpretation of the model, and it can also be supported when agents can always find people to question about their periodograms who have sufficiently long income histories. A natural question, though, is how our results are changed when the cost of acquiring information varies across frequencies.

In this subsection, we consider the following alternative to assumption 4:

$$\sum_{j=1}^{n} \phi\left(j\right) \tau\left(\omega_j\right) d\omega \leq \bar{\tau} \tag{36}$$

for some cost function ϕ. It does not appear possible to obtain a closed-form solution for optimal attention, τ^*, under general ϕ. However, in the special case where there is no smoothing across frequencies – $\lambda = 0$ – it does remain possible to find a solution:

$$\text{for } \lambda = 0,\ \tau^*\left(\omega_j\right) = Z\left(\omega_j\right) \phi\left(\omega_j\right)^{-1/2} \theta^{-1/2} \psi^{1/2}. \tag{37}$$

(37) is a simple generalization of the result in the baseline case; the only difference is that now

[23] See Cochrane (1989) for an more extended analysis of this issue and Eichenbaum (2011) for discussion in a setting closely related to ours.

$\tau^*(\omega_j)$ is decreasing in the cost of obtaining information at frequency ω_j. If low frequencies are more expensive to learn about than higher frequencies, then τ^* will have a less extreme tilt toward low frequencies than in the baseline case.

Recall our motivation for the learning framework in which agents get information about the dynamics of income by asking other people. In order to have information about a particular frequency ω, the person asked must have been alive for at least $2\pi/\omega$ periods (that is the first periodogram ordinate; intuitively, one does not have any direct information about fluctuations that last longer than the data sample). So if only some fraction $F(\omega)$ of people have been alive for at least $2\pi/\omega$ periods, then on average an agent must talk to $1/F(\omega)$ people in order to find a person who can inform them about frequency ω.

More specifically, suppose people die with a probability $\delta \in (0,1)$ in every period. Then as long as the birth rate is constant, the fraction of people who have been alive for at least k periods is δ^{k-1}, implying that $F(\omega) = \delta^{(2\pi/\omega)-1}$. A reasonable functional form for ϕ is therefore

$$\phi(\omega) = \delta^{1-(2\pi/\omega)}. \tag{38}$$

As $\omega \to 0$, $\phi(\omega) \to \infty$, which means that in general this cost function will cause agents to learn less about low frequencies than in the baseline. However, $\phi'(0) = -\infty$, while $Z'(0) = 0$. So attention should be increasing with frequency local to zero.

We calibrate $\delta = 0.975$, corresponding to an annual death probability of 2 percent, which we motivate as equivalent to people having a 50-year working life on average. The top panels of figure 7 then plot the optimal information policies $\tau^* \propto Z$ and $\tau^\phi \propto Z\phi^{-1/2}$ (normalized to have equal integrals). Both lines again peak at low frequencies, but whereas τ^* peaks at frequency zero, τ^ϕ peaks at a slightly interior frequency. That peak comes at a frequency corresponding to cycles lasting approximately 160 years, though. So while the function ϕ causes agents to learn less about the very lowest frequencies, they still very much focus their attention on long-term cycles.

To see how that change affects our calibration, the bottom panels of figure 7 plot the worst-case spectra under various τ policies now also including $\tau^\phi(\omega)$.[24] That policy leads to results between the benchmark τ^* and the constant τ policy. At the very lowest frequencies, the τ^ϕ model does not match the true spectrum as well as τ^*, but it still does much better than τ^F. At the middle frequency peak and at higher frequencies, on the other hand, the policy τ^ϕ does a better job of matching the log spectrum than τ^* but still worse than τ^F.

This section therefore shows, as one might expect, that when low frequencies are more costly to learn about, the main results are weakened somewhat. We continue to find that agents allocate the most attention to low frequencies, just not to the *very* lowest – the peak is at an interior frequency, but one corresponding to cycles lasting a century or more. The impact on the worst-case model is

[24] For non-zero λ with ϕ varying across frequencies, $\tau^\phi(\omega) \propto Z(\omega)\phi(\omega)^{-1/2}$ is not technically the optimal policy – it must be solved for numerically. We focus on the analytic case for the sake of simplicity. Furthermore, the calibration in figure 7 is set up so that the total precision under τ^* is the same as that under τ^ϕ – they differ only in how that precision is allocated across frequencies.

to put it somewhere between that induced in the baseline optimum and that induced by the policy that puts equal weight on all frequencies.

7 Conclusion

This paper studies how people can direct their attention to different features of a model. We consider a nonparametric class of income processes and show precisely how agents optimally allocate attention to the behavior of income at different frequencies. The utility maximizing policy is to pay the most attention to the behavior of income at very low frequencies, and use a relatively simple and inaccurate model at high frequencies.

While there is extensive past work on learning, the innovation of this paper is to provide an exactly solvable framework for studying how learning can be applied to different aspects of a model of the world, as opposed to learning about state variables. The theory can be used to describe what people pay attention to, what aspects of the world they try to model accurately and what they use coarser approximations for, and the set of mistakes that people should be expected to make.

We show that optimal learning implies people are most likely to make mistakes at high frequencies, as those are the aspects of the income process least important for utility. Consistent with empirical evidence, the model implies that consumption tends to track transitory fluctuations in income in the short-run, but at lower frequencies consumption growth is close to white noise (which it would be under the full-information optimal policy). In other words, the consumption mistakes that the empirical literature has documented are consistent with optimal learning.

References

Abel, Andrew B, Janice C Eberly, and Stavros Panageas, "Optimal inattention to the stock market," *The American economic review*, 2007, *97* (2), 244–249.

____ , ____ , and ____ , "Optimal inattention to the stock market with information costs and transactions costs," *Econometrica*, 2013, *81* (4), 1455–1481.

Akaike, Hirotugu, "Smoothness Priors and the Distributed Lag Estimator.," Technical Report, DTIC Document 1979.

Bansal, Ravi and Ivan Shaliastovich, "Confidence Risk and Asset Prices," *The American Economic Review*, 2010, *100* (2), 537–541.

Barron, John M and Jinlan Ni, "Endogenous asymmetric information and international equity home bias: the effects of portfolio size and information costs," *Journal of International Money and Finance*, 2008, *27* (4), 617–635.

Brillinger, David R., *Time Series: Data Analysis and Theory*, McGraw Hill, 1981.

Chow, Yun-Shyong and Ulf Grenander, "A Sieve Method for the Spectral Density," *The Annals of Statistics*, 1985, *13(3)*, 998–1010.

Christiano, Lawrence J., "Notes on Fuster, Hebert and Laibson, "Natural Expectations, Macroeconomic Dynamics, and Asset Pricing"," 2011. Lecture notes downloaded from http://faculty.wcas.northwestern.edu/~lchrist/finc520/note_on_fhl.pdf.

Cochrane, John H, "The Sensitivity of Tests of the Intertemporal Allocation of Consumption to Near-Rational Alternatives," *The American Economic Review*, 1989, *79* (3), 319–337.

Collin-Dufresne, Pierre, Michael Johannes, and Lars Lochstoer, "Parameter Learning in General Equilibrium: The Asset Pricing Implications," *American Economic Review*, Forthcoming.

Dew-Becker, Ian, "How risky is consumption in the long-run? Benchmark estimates from a novel unbiased and efficient estimator," 2015. Working paper.

___ **and Stefano Giglio**, "Asset Pricing in the Frequency Domain: Theory and Empirics," *Review of Financial Studies*, 2016. Forthcoming.

Diaconis, Persi and David Freedman, "On the consistency of Bayes estimates," *The Annals of Statistics*, 1986, *14* (1), 1–26.

Eichenbaum, Martin, "Comment on "Natural Expectations, Macroeconomic Dynamics, and Asset Pricing"," in "NBER Macroeconomics Annual 2011, Volume 26," University of Chicago Press, 2011, pp. 49–60.

Ellsberg, Daniel, "Risk, ambiguity, and the Savage axioms," *The Quarterly Journal of Economics*, 1961, *75* (4), 643–669.

Fuster, Andreas, Benjamin Hebert, and David Laibson, "Natural Expectations, Macroeconomic Dynamics, and Asset Pricing," *NBER Macroeconomics Annual*, 2011, *26* (1), 1–48.

Gabaix, Xavier, "A Sparsity-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 2014, *129* (4), 1661–1710.

___ , "Behavioral Macroeconomics Via Sparse Dynamic Programming." Working paper.

Gersch, Will and Genshiro Kitagawa, "Smoothness priors transfer function estimation," *Automatica*, 1989, *25* (4), 603–608.

Gigerenzer, Gerd, "Fast and frugal heuristics: The tools of bounded rationality," *Blackwell handbook of judgment and decision making*, 2004, pp. 62–88.

Gilboa, Itzhak and David Schmeidler, "Maxmin expected utility with non-unique prior," *Journal of mathematical economics*, 1989, *18* (2), 141–153.

Hansen, Lars P. and Thomas J. Sargent, *Robustness*, Princeton University Press, 2007.

___ and ___ , "Fragile Beliefs and the Price of Uncertainty," *Quantitative Economics*, 2010, *1(1)*, 129–162.

Hansen, Lars Peter and Thomas J Sargent, "Formulating and estimating dynamic linear rational expectations models," *Journal of Economic Dynamics and control*, 1980, *2*, 7–46.

___ and ___ , "A note on Wiener-Kolmogorov prediction formulas for rational expectations models," *Economics Letters*, 1981, *8* (3), 255–260.

___ and Thomas J. Sargent, "Sets of Models and Prices of Uncertainty," 2015. Working paper.

___ , Thomas J Sargent, and Thomas D Tallarini, "Robust permanent income and pricing," *Review of Economic studies*, 1999, *66* (4), 873–907.

Hsieh, Chang-Tai, "Do Consumers React to Anticipated Income Changes? Evidence from the Alaska Permanent Fund," *The American Economic Review*, 2003, *93* (1), 397–405.

Jappelli, Tullio and Luigi Pistaferri, "The Consumption Response to Income Changes," *Annu. Rev. Econ*, 2010, *2*, 479–506.

Ju, Nengjiu and Jianjun Miao, "Ambiguity, Learning, and Asset Returns," *Econometrica*, 2012, *80(2)*, 559–591.

Kacperczyk, Marcin, Stijn Van Nieuwerburgh, and Laura Veldkamp, "Rational attention allocation over the business cycle," Technical Report, National Bureau of Economic Research 2016.

Kaplan, Greg and Giovanni L Violante, "A Model of the Consumption Response to Fiscal Stimulus Payments," *Econometrica*, 2014, *82* (4), 1199–1239.

Kitagawa, Genshiro and Will Gersch, "A smoothness priors–state space modeling of time series with trend and seasonality," *Journal of the American Statistical Association*, 1984, *79* (386), 378–389.

___ and ___ , "A smoothness priors long AR model method for spectral estimation," *IEEE transactions on automatic control*, 1985, *30* (1), 57–65.

___ and ___ , *Smoothness Priors Analysis of Time Series*, Springer Science & Business Media, 1996.

Knight, Frank H., *Risk, Uncertainty, and Profit*, Houghton Mifflin, 1921.

Kueng, Lorenz, "Explaining Consumption Excess Sensitivity with Near-Rationality: Evidence from Large Predetermined Payments," 2016. Working paper.

Lipman, Barton L, "Information processing and bounded rationality: a survey," *Canadian Journal of Economics*, 1995, pp. 42–67.

Luo, Yulei, "Consumption dynamics under information processing constraints," *Review of Economic dynamics*, 2008, *11* (2), 366–385.

___ **and Eric R Young**, "Risk-sensitive consumption and savings under rational inattention," *American Economic Journal: Macroeconomics*, 2010, *2* (4), 281–325.

Machina, Mark J. and Marciano Siniscalchi, "Ambiguity and Ambiguity Aversion," in "Handbook of the Economics of Risk and Uncertainty," Elsevier, 2014, pp. 729–807.

Nieuwerburgh, Stijn Van and Laura Veldkamp, "Information acquisition and under-diversification," *The Review of Economic Studies*, 2010, *77* (2), 779–805.

Optimal Sticky Prices under Rational Inattention, *The American Economic Review*, 2009, *99* (3), 769–803.

Parker, Jonathan A, "The Reaction of Household Consumption to Predictable Changes in Social Security Taxes," *The American Economic Review*, 1999, *89* (4), 959–973.

Payne, John W and James R Bettman, "Walking with the scarecrow: The information-processing approach to decision research," *Blackwell handbook of judgment and decision making*, 2004, pp. 110–132.

Peng, Lin and Wei Xiong, "Investor attention, overconfidence and category learning," *Journal of Financial Economics*, 2006, *80* (3), 563–602.

Priestley, M.B., *Spectral Analysis and Time Series*, Elsevier, 1981.

Reis, Ricardo, "Inattentive consumers," *Journal of monetary Economics*, 2006, *53* (8), 1761–1800.

Schwartzstein, Joshua, "Selective attention and learning," *Journal of the European Economic Association*, 2014, *12* (6), 1423–1452.

Shiller, Robert J, "A distributed lag estimator derived from smoothness priors," *Econometrica: journal of the Econometric Society*, 1973, pp. 775–788.

Sims, Christopher A, "Distributed lag estimation when the parameter space is explicitly infinite-dimensional," *The Annals of Mathematical Statistics*, 1971, *42* (5), 1622–1636.

___ , "Implications of Rational Inattention," *Journal of Monetary Economics*, 2003, *50* (3), 665–690.

Souleles, Nicholas S, "The Response of Household Consumption to Income Tax Refunds," *The American Economic Review*, 1999, *89* (4), 947–958.

van Nieuwerburgh, Stijn and Laura Veldkamp, "Information acquisition and under-diversification," *The Review of Economic Studies*, 2010, *77* (2), 779–805.

Veldkamp, Laura L, "Information markets and the comovement of asset prices," *The Review of Economic Studies*, 2006, *73* (3), 823–845.

Veldkamp, Laura L., *Information Choice in Macroeconomics and Finance*, Princeton University Press, 2011.

Wahba, Grace, "Automatic smoothing of the log periodogram," *Journal of the American Statistical Association*, 1980, *75* (369), 122–132.

Wang, Neng, "Precautionary saving and partially observed income," *Journal of Monetary Economics*, 2004, *51* (8), 1645–1681.

Whittle, Peter, *Hypothesis Testing in Time Series Analysis*, Vol. 4, Almqvist and Wiksells, 1951.

Figure 1: Weighting function $Z(\omega)$ and its multiplicative inverse

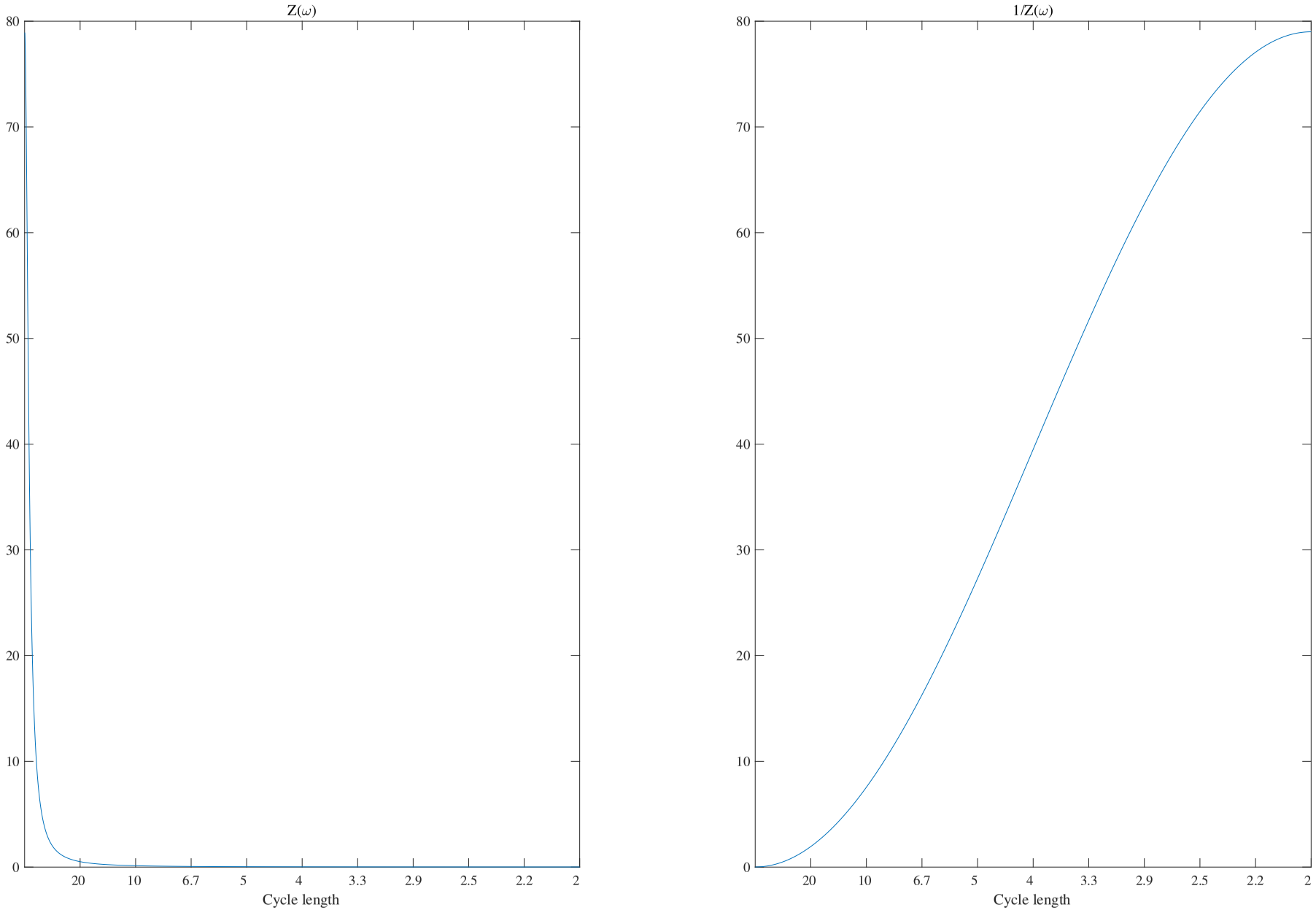


Figure 2: Average estimated log spectra and IRFs

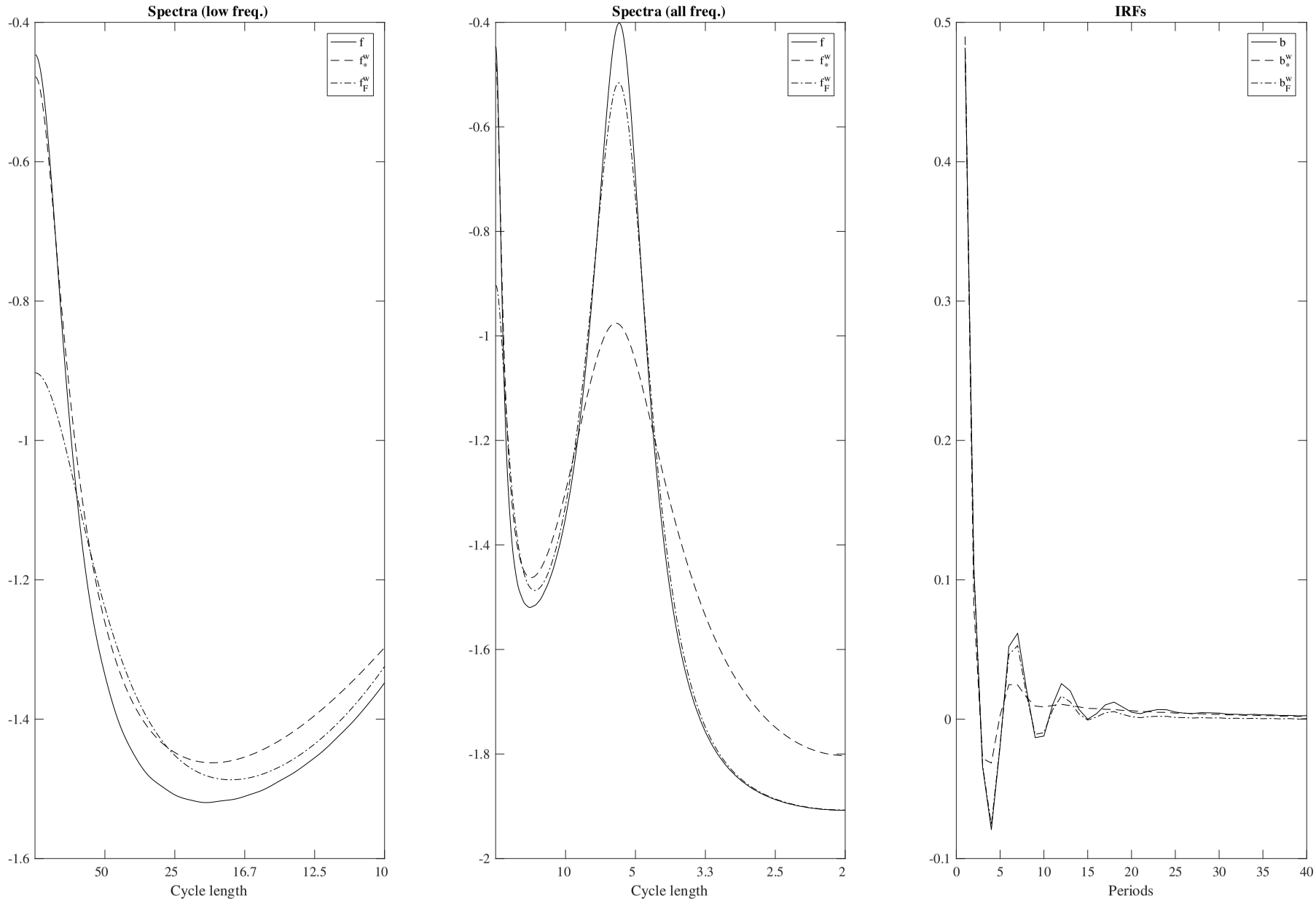
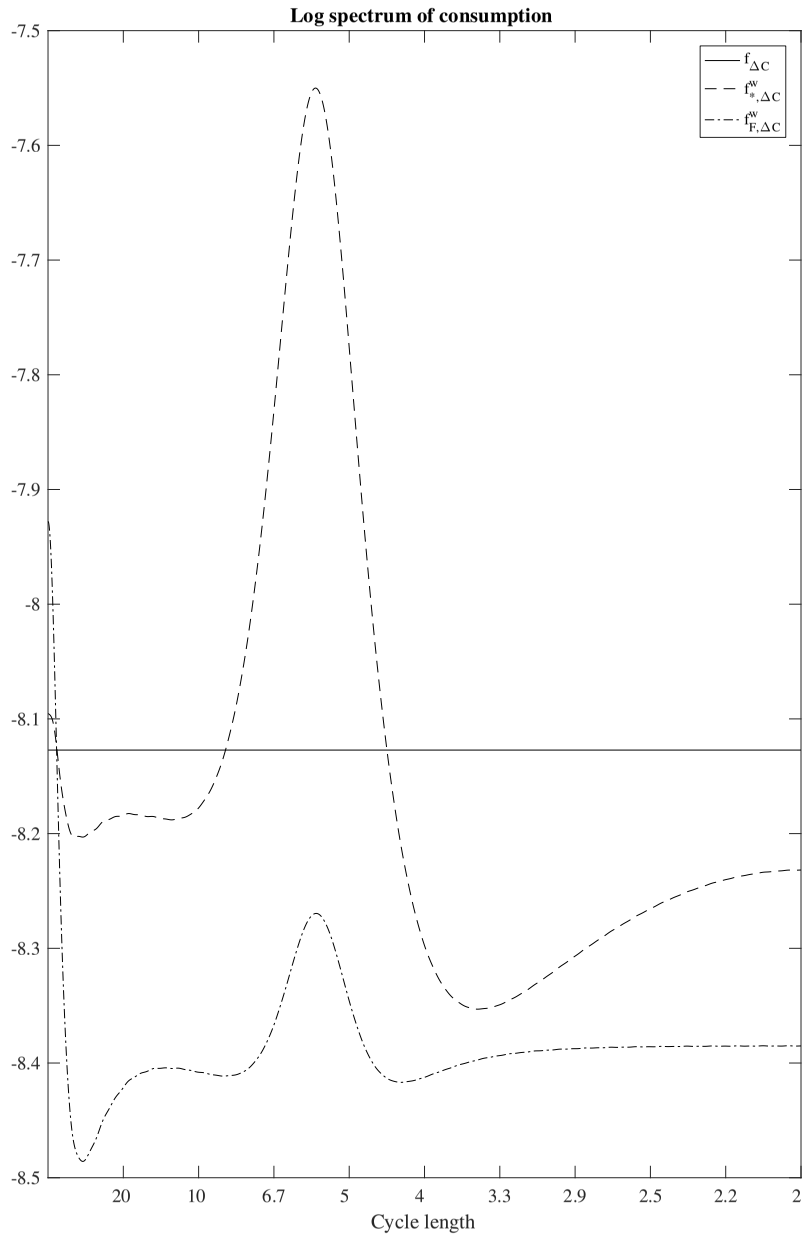


Figure 3: Behavior of consumption



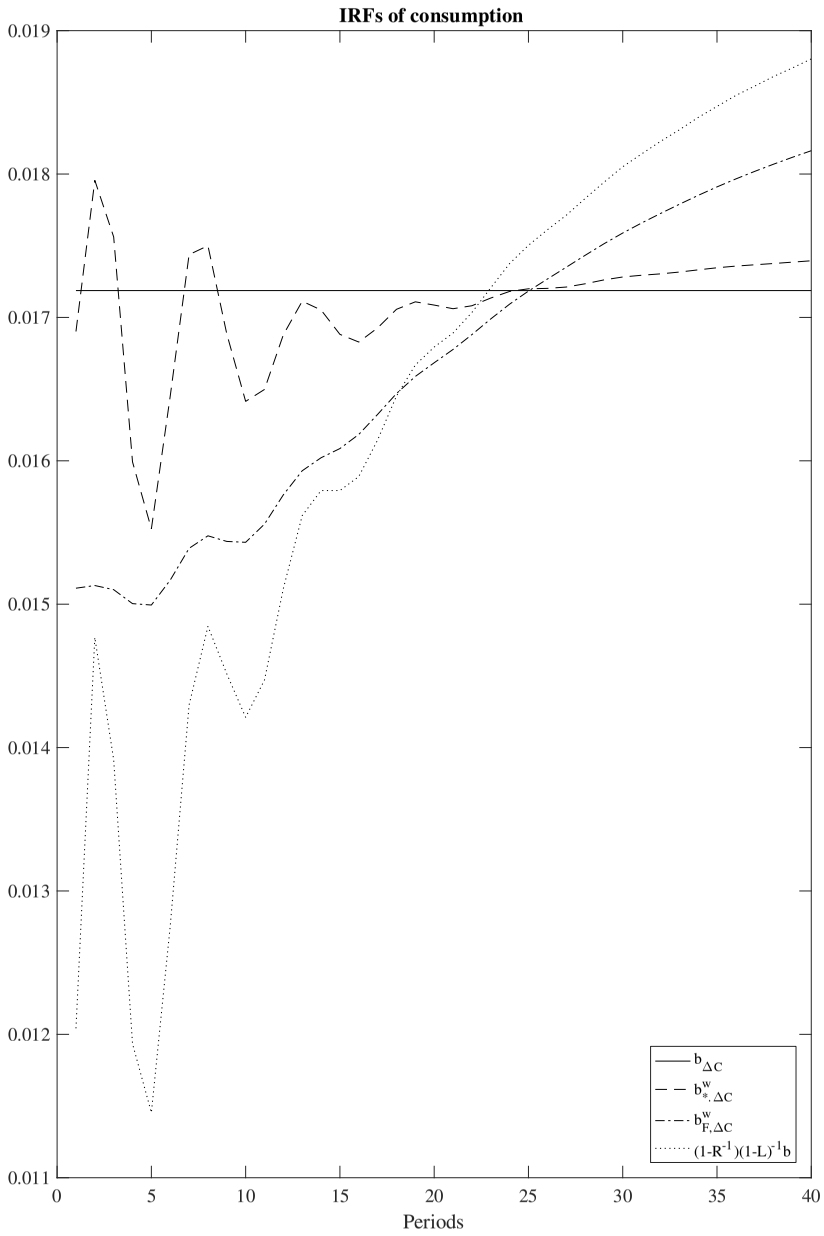
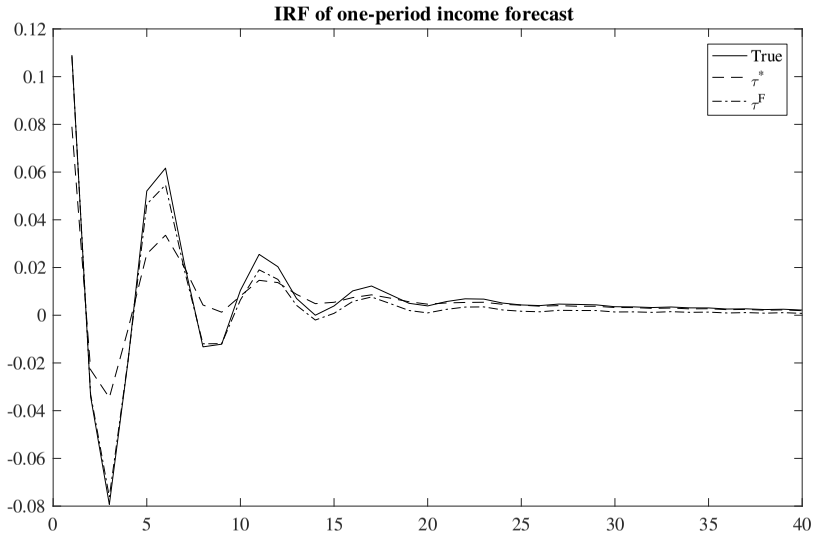
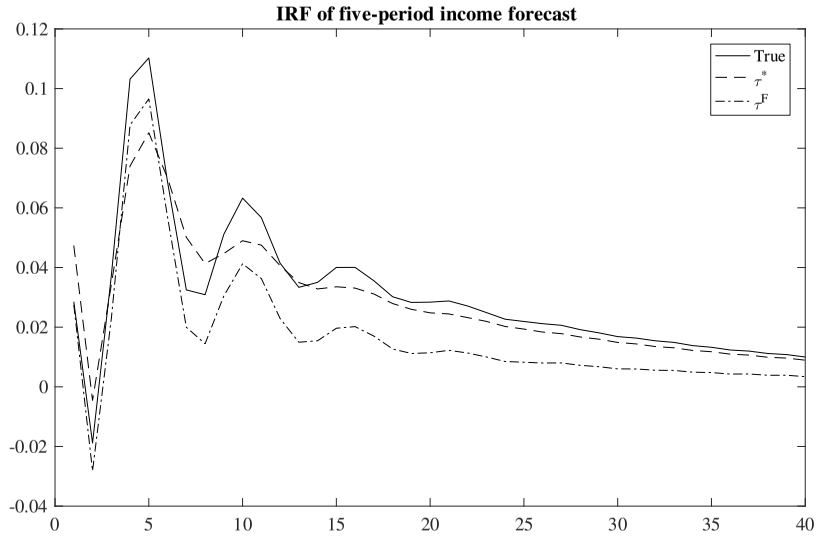
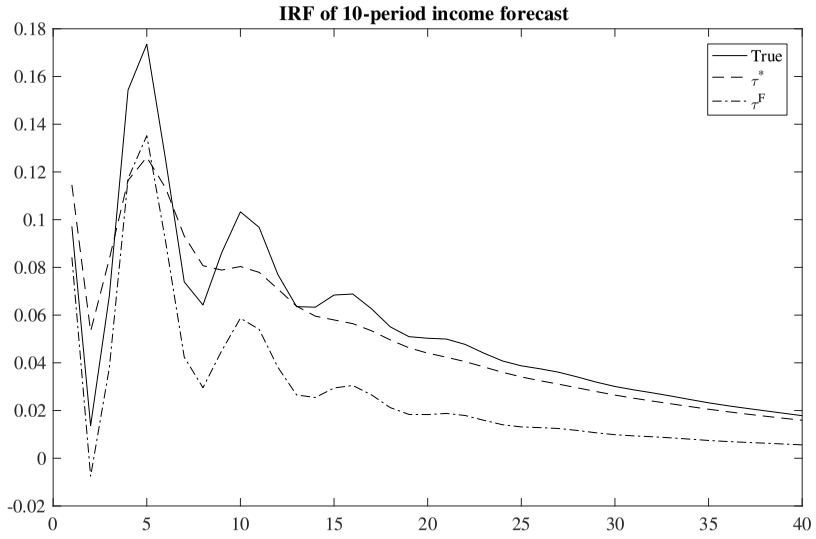


Figure 4: IRFs of income forecasts







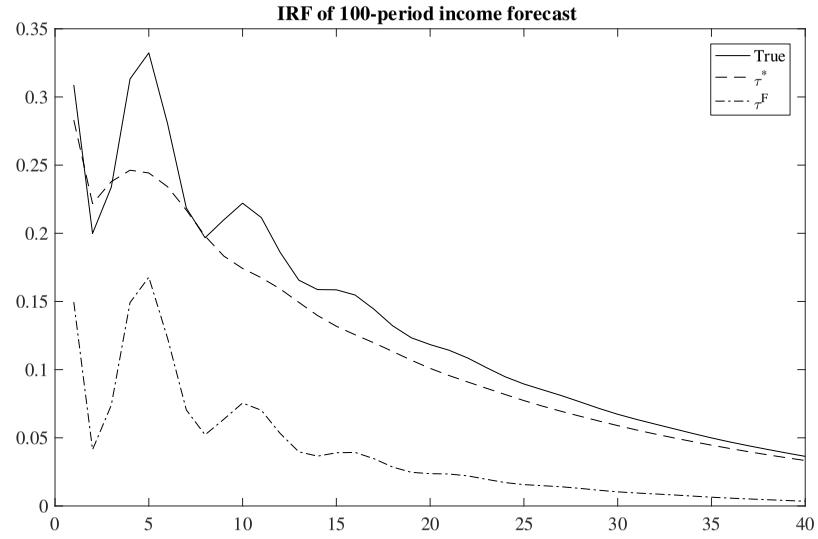


Figure 5: Return forecasting R^2's

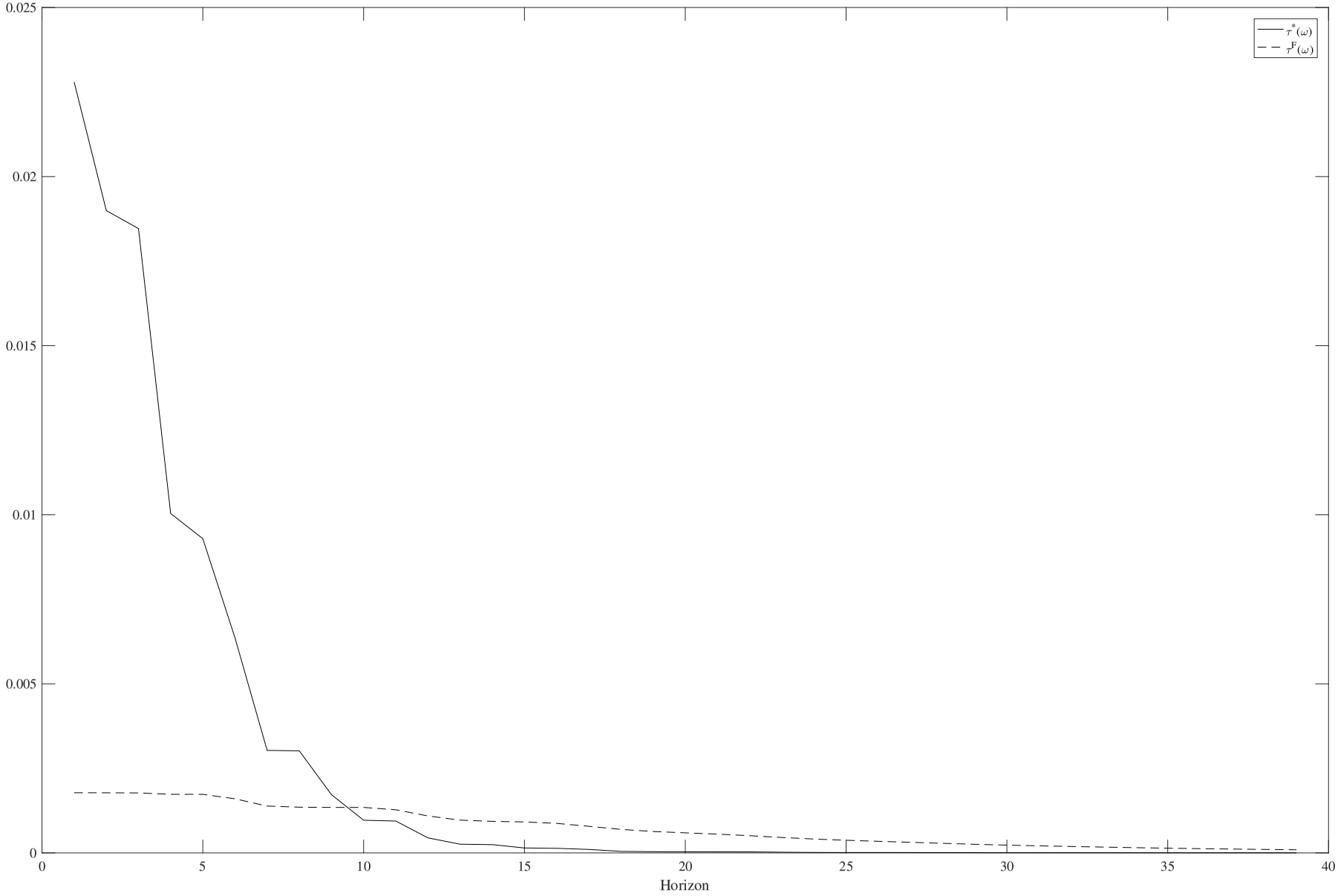
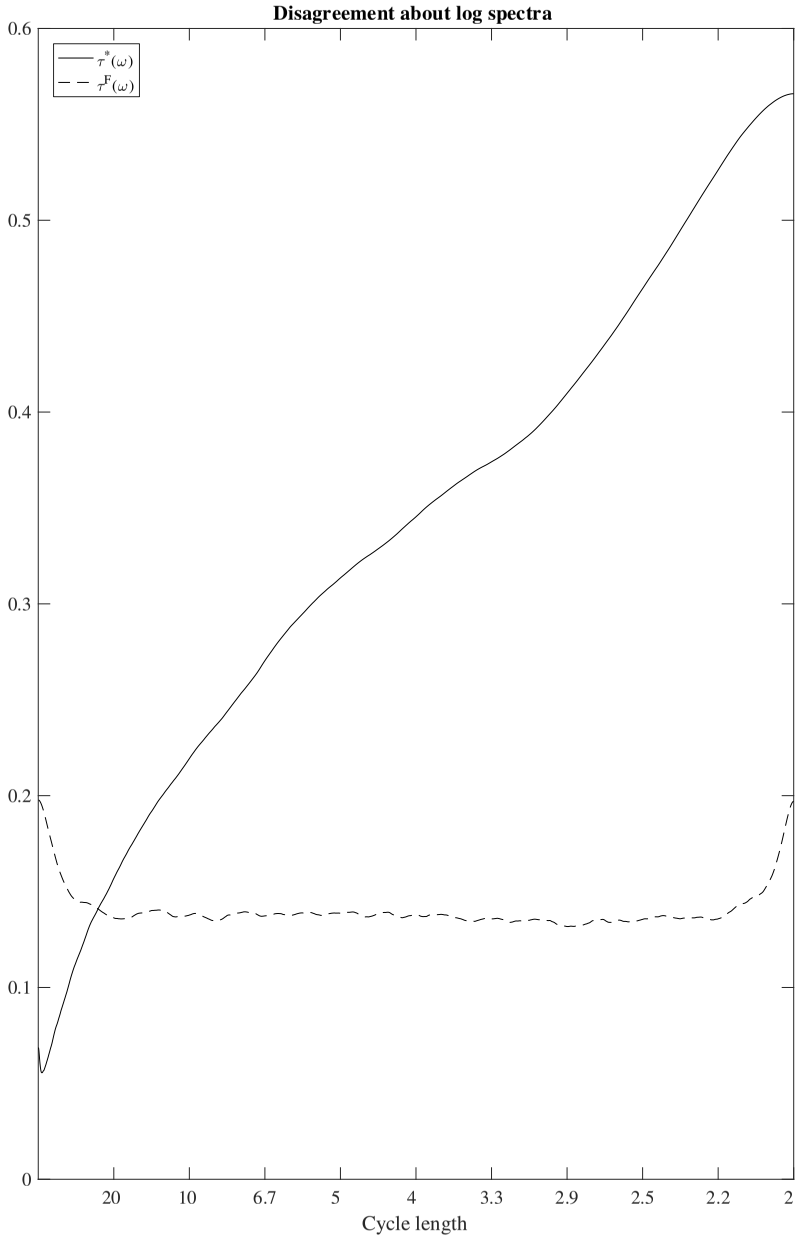


Figure 6: Disagreement



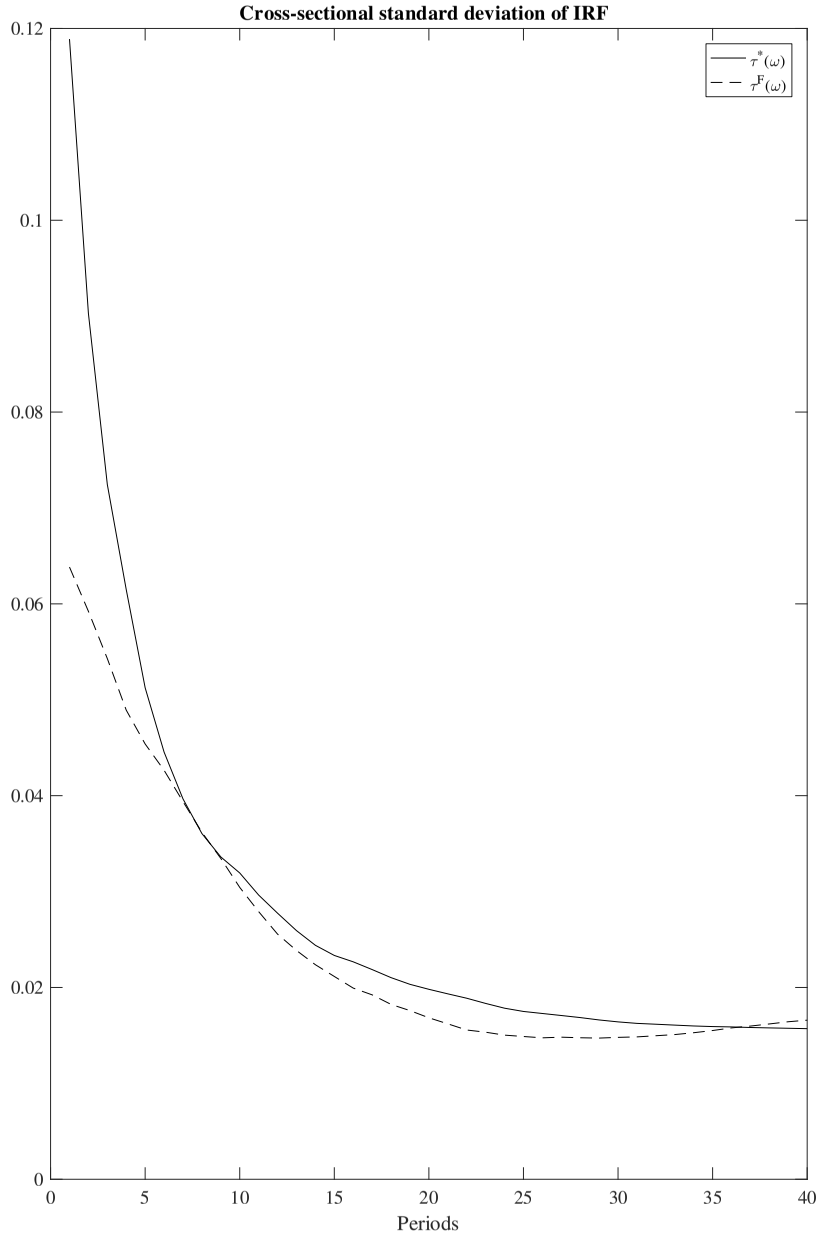
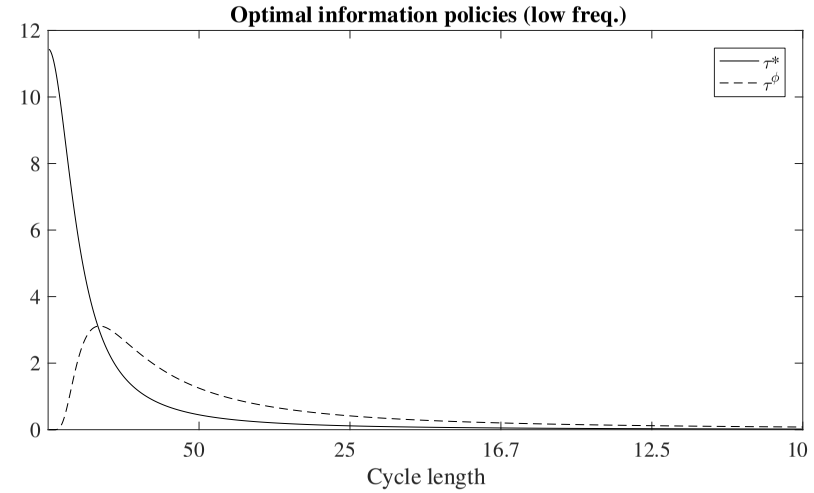
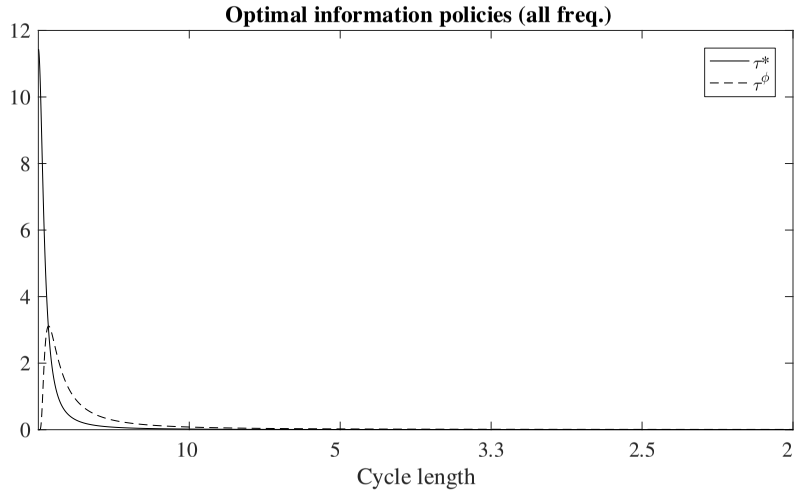
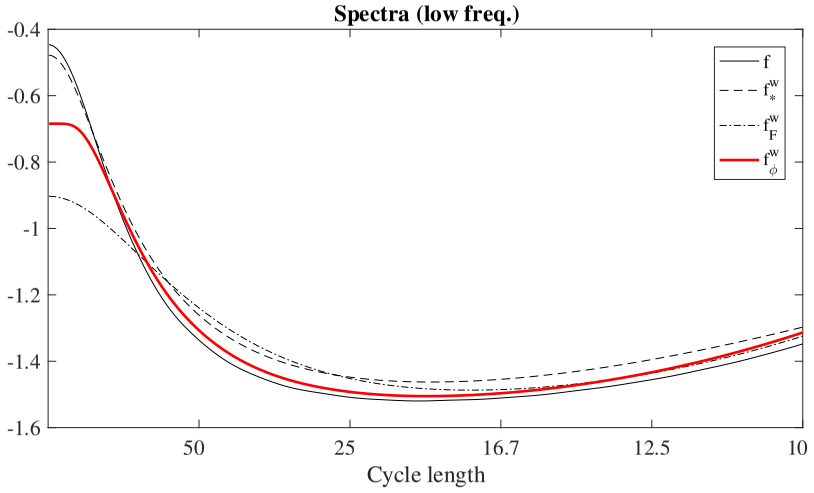


Figure 7: Effects of information cost varying across frequencies







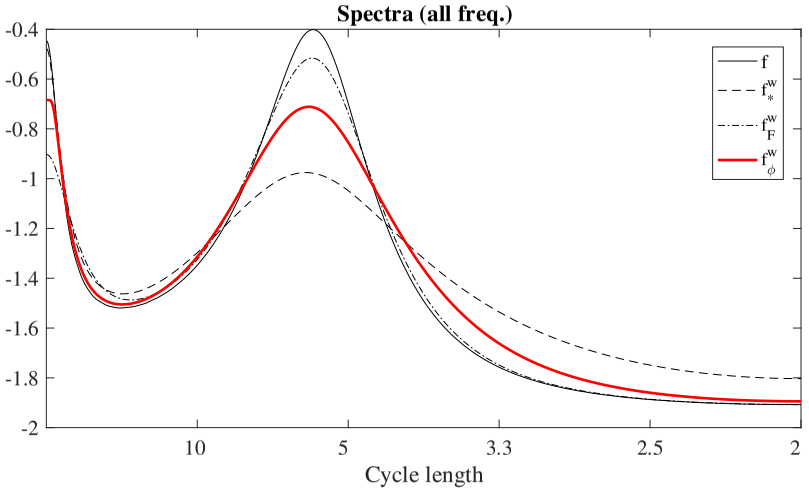


Figure 8: IRFs of cumulative asset returns

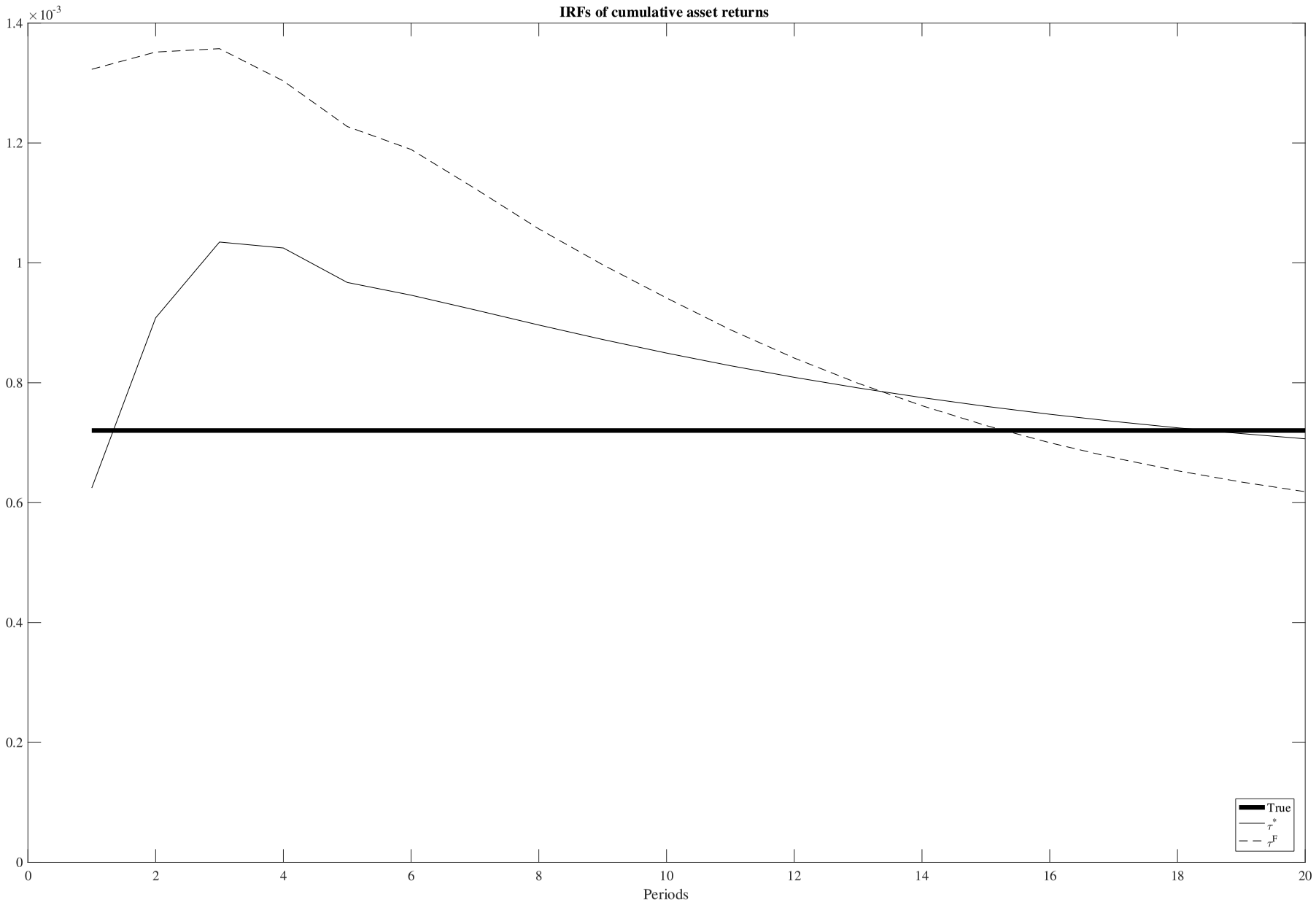


Figure 9: Average estimated log spectra and IRFs with momentum and mean reversion