Directed attention and nonparametric learning

Ian Dew-Becker and Charles G. Nathanson*

March 10, 2019

Abstract

This paper examines the implications of learning for the effects of ambiguity aversion. The key result is that since agents naturally choose to learn about the sources of uncertainty that reduce utility the most, information acquisition attenuates the most severe effects of ambiguity aversion. The specific setting we study is the canonical consumption/savings problem. Agents endogenously learn most about income dynamics at the very lowest frequencies. While ambiguity aversion typically implies in this setting excessive extrapolation of income shocks, that effect is eliminated here. Furthermore, deviations of consumption from the full-information benchmark are largest at high frequencies, so the model naturally generates overreaction of consumption to predictable short-run income variation.

A large recent literature studies model uncertainty, and ambiguity aversion in particular, as a major driver of macroeconomic dynamics and asset prices. Ambiguity aversion has been shown to be able to generate realistic business cycles (Ilut and Schneider (2014) and Bianchi, Ilut, and Schneider (2017)), to help rationalize variation in survey expectations (Bhandari, Borovicka, and Ho (2017)), to generate large and time-varying equity risk premia (Hansen, Sargent, and Tallarini (1999), Ju and Miao (2012), Hansen and Sargent (2015), and Bidder and Dew-Becker (2016)) and to help explain the VIX and the variance risk premium (Drechsler (2013) and Bidder and Smith (2015)).

The central idea behind the ambiguity aversion literature is that people are uncertain about the true model driving the economy and that they choose policies that are designed to be robust against unfavorable models. While the ambiguity literature has taken the model uncertainty to be exogenous, one would naturally expect that if people were highly averse to model uncertainty that they would try to learn and reduce that ambiguity. And that learning should be focused on precisely the parts of the underlying model where errors are most painful. So learning should be expected to reduce the most severe effects of ambiguity aversion and model uncertainty.

This paper studies ambiguity and learning in a simple dynamic consumption/savings problem. The basic intuition above could be captured in many settings, but we choose the consumption/savings problem because it is a canonical dynamic optimization with well understood analytic

*Dew-Becker: Northwestern University and NBER; ian.dewbecker@gmail.com. Nathanson: Northwestern University; nathanson@kellogg.northwestern.edu. We appreciate helpful comments from Ben Hebert, Peter Klibanoff, Konstantin Milbradt, Mikkel Plagborg-Møller, and seminar participants.

solutions that shares the same basic structure as much richer models used in asset pricing and macroeconomics.[1]

The agents in the model face an exogenous income process with uninsurable risk and unknown dynamics, and they are ambiguity averse over potential models for income. This paper's key innovation compared to past work is that agents acquire information that can reduce the degree of ambiguity, and that acquisition can be directed to different aspects of the income process. Our goal is to find the optimal information acquisition policy and understand how it changes the effects of ambiguity on an agent's behavior. The key analytic result is that there is a simple benchmark case in which optimal learning completely eliminates the primary effects of ambiguity aversion, regardless of the total quantity of information acquired. More generally, learning always acts to reduce the effects, with the degree depending on the details of the cost specification for information.

The optimization problem that agents face has three components: the consumption choice, nature's choice of a model (embodying ambiguity aversion), and learning. Conditional on a particular model of the world, agents have standard Bayesian expected utility.[2] The agents are unsure of the true model, though, which is where their ambiguity aversion appears: agents act as though nature chooses the process for income, among all sufficiently plausible processes, that will yield the lowest utility from consumption. This selection criterion ensures that consumption decisions are robust to uncertainty about the true model.

The third phase of the optimization represents our contribution. Agents allocate attention to different aspects of the income process, which allows them to endogenously limit the degree of ambiguity they face. When agents pay more attention to a particular aspect of income, such as its low-frequency behavior, they receive information about its true behavior along that dimension and the set of plausible models narrows. A contribution of the paper to the learning literature is in providing a general description of how a person might learn about different aspects of a dynamic process.[3] Kasa (2006) provides a related analysis of the link between ambiguity and information acquisition, but the analysis applies to the total quantity of information acquired, whereas the key mechanism here is the choice of what to learn about.

Given that optimal consumption depends on permanent income, it is the low-frequency features of income that are generally most beneficial to learn about. But one must also ask how costly it is to learn about dynamics at different frequencies. Textbook results from the time series econometrics literature say that learning about all frequencies is equally hard (e.g. Brillinger (1981), Priestly (1981), Brockwell and Davis (1991), and Hamilton (1994)). But since intuition suggests that low

[1] See Wang (2004, 2009) and Luo (2008) for analyses of consumption under model uncertainty and information processing constraints and Caballero (1990) for an analysis of the setup with a known model.

[2] During this phase of the optimization, no dynamic learning about the model occurs. For boundedly rational models of dynamic learning, see 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).

[3] 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. See Sims (2003), Veldkamp (2011), and many citations therein for work on directed attention more generally.

frequencies might be more difficult to learn about, we also consider a general specification that allows for arbitrary costs across frequencies.

The full optimization is analytically tractable which allows us to sharply establish our main result: when information is equally costly across frequencies, the agent directs almost all attention to the behavior of income at the lowest frequencies (i.e. at long horizons), which have the largest impact on utility through their ambiguity aversion. The agent's learning in that case perfectly cancels out the most harmful effects of ambiguity. More specifically, we obtain two key results:

1. In past work – in which agents have no ability to acquire information – ambiguity aversion causes agents to generally overextrapolate income shocks (Hansen and Sargent (2010, 2017) and Bidder and Dew-Becker (2016)),[4] but with endogenous learning, that result is completely eliminated: agents neither over- nor under-extrapolate shocks when forecasting long-run future income. The lack of bias results from the fact that agents acquire the most information at low frequencies (regardless of the total quantity acquired).

2. The agents' focus on low frequencies yields high-frequency mistakes: at short horizons, consumption growth is positively correlated with the *predictable* component of income growth. This comovement violates the permanent income hypothesis (Friedman (1957), Hall (1978)) but matches the extensive empirical evidence on the excess sensitivity of consumption to income (Jappelli and Pistaferri (2010), Kaplan and Violante (2014)). Because agents fail to learn about the high-frequency characteristics of the income process, much of the predictable variation in income is surprising and therefore leads agents to adjust consumption.

What connects the two theoretical results is that high-frequency mistakes have minimal implications for lifetime utility, while low-frequency mistakes can have substantial effects. That idea has been suggested as an explanation for the excess sensitivity puzzle, and the present model formalizes it.[5] People cannot achieve perfection, so they choose to make mistakes that are minimally costly. Two aspects of the results are surprising. First, while one might expect that learning would reduce the effects of ambiguity (though that point has not been made previously), the fact that it can perfectly cancel those effects in some cases – even though the learning is incomplete in the sense that not all uncertainty is resolved – has important implications for the interpretation of ambiguity models. Second, while it is understood that the utility cost of high-frequency mistakes is relatively small, this is the first paper to obtain such mistakes endogenously, and we show that they can be

[4] A bias towards belief in overly persistent processes is present also in the boundedly rational frameworks of Fuster, Hebert, and Laibson (2011) and Bordalo, Gennaioli, and Shleifer (2016). Beyond the fact that this paper allows for information acquisition, it also differs from Hansen and Sargent (2010, 2017) and Bidder and Dew-Becker (2016) in that there is an endogenous consumption/savings decision. The fact that low-frequency fluctuations are most important here is thus an endogenous result (rather than assumed through the use of generalized recursive preferences). The learning, moreover, completely eliminates the excessive extrapolation that is the main result in Bidder and Dew-Becker (2016).

[5] See Cochrane (1989), Eichenbaum (2011), and Kueng (2016) for discussions of the small utility costs of excess sensitity to transitory income shocks.

quantitatively realistic. The paper is thus important both to the literature on ambiguity aversion in dynamic models and also the literature on "mistakes" in household consumption.

The main findings hold in the textbook benchmark where information is equally costly at all frequencies. We also formalize the idea that low-frequency information should be more expensive to obtain than high frequency information. In that case, the agent continues to focus primarily on low frequencies, but with a less extreme tilt. Weighted by the precision of the signals, the median unit of attention is focused on cycles lasting 250 years in the benchmark case (consistent with results in Dew-Becker and Giglio (2014)) and 47 years in the frequency-dependent cost case. So while attention shifts to much shorter cycles, it is still focused on extremely long-lived shocks. In terms of observable behavior, the results in this case lie between the equal information benchmark and the case of ambiguity with no attention allocation. The bottom line of the analysis, then, is that the effects of ambiguity aversion depend critically on how easily agents can acquire information. In a reasonable benchmark, the main effects can be completely eliminated, but there are also specifications for information costs that generate intermediate outcomes.

In addition to the work on ambiguity aversion above, our work links to the literature on learning and attention allocation more generally (e.g. Sims (2003)). Most past work has focused on learning about hidden states (e.g. Guvenen (2007)). A potentially useful contribution of the paper is to propose a framework for analyzing information acquisition about specific aspects of a dynamic process and for motivating how the cost might vary across different features.[6] Contemporaneous work by Epstein and Ji (2017) also examines learning under ambiguity, but in a setting in which there is no choice about how to allocate attention and in which the source of ambiguity is not dynamic.

The result that the effects of ambiguity are eliminated by learning presumes that learning is possible, and here it is relatively straightforward since income is stationary. An implication of the results is that for ambiguity to have major effects, dynamic models may need to incorporate nonstationarity or regime shifts, which would make learning much more difficult, and also more realistic.

1 The optimization problem

We study a consumption/savings problem with endogenous information acquisition. The consumption policy choice is affected by the fact that agents are unsure of the true process driving income. Agents are ambiguity averse and choose their consumption rule to be robust to an unfavorable income process. The consumption choice and ambiguity aversion are relatively standard. What is novel is that the model endogenizes the set of income processes over which agents are ambiguity averse.

While some of the ideas could be illustrated in a simpler model, there are a number of factors that lead to our choice of the setting to study. First, specifications closely related to this setup have

[6] See Gabaix (2016) for a recent alternative model of directed attention in a dynamic setting.

been studied in a number of papers in the ambiguity literature discussed above, so it is valuable to understand how the results in those models are affected by information acquisition. Second, the fact that the model has an infinite horizon means that all autocorrelations in income affect utility, giving the agent a rich space from which to choose information policies. We are able to derive general results on which autocorrelations have the strongest effect on utility that apply in canonical consumption/savings models. The infinite horizon also allows us to take advantage of powerful nonparametric methods that are not available in finite settings. Finally, the information allocation choice in the model is consequential because it has direct and observable effects on the persistence of consumption. Much has been made of the question of whether consumption is a random walk, and this model has direct implications for that.

This section lays out the basic optimization problem, and the information acquisition follows subsequently.

1.1 Consumption and ambiguity aversion

1.1.1 Income

Agents face a standard budget constraint and income process.

Assumption 1 *Financial wealth, W_t, follows the process*

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

where C_t is consumption, Y_t is an exogenous income stream, and R is a fixed gross interest rate. Income follows

$$\begin{aligned} Y_t &= a(L) Y_{t-1} + b_0 \varepsilon_t & (2)\\ \varepsilon_t &\sim \textit{i.i.d. } N(0,1) & (3) \end{aligned}$$

where $a(L)$ is a power series in the lag operator, L (where $L^j x_t = x_{t-j}$). We assume $a(L)$ is such that Y is well behaved (in particular, has a spectrum that is positive and bounded). The ε_t are unobservable.

The assumptions of linearity and Gaussianity are in line with past work, but can be relaxed – Gaussianity is not necessary, for example.[7] Most of the analysis uses the Wold representation,

$$\begin{aligned} Y_t &= b(L)\varepsilon_t, & (4)\\ \text{where } b(L) &\equiv \frac{b_0}{1 - La(L)}. & (5) \end{aligned}$$

[7] The critical assumptions are that income is second-order stationary and that it has a spectral density that is finite and bounded away from zero. The distribution of the innovations is largely irrelevant beyond existence conditions, but it is important that it is fixed over time.

The coefficients in the power series $b(L)$ are denoted b_j (i.e. $b(L) = \sum_{j=0}^{\infty} b_j L^j$). A convenient feature of the representation b is that the coefficient b_j represents the impulse response of Y_{t+j} to ε_t, so that the b_j's trace out the full impulse response function.

Throughout the paper, we refer to models in the time domain in terms of $b(L)$. Since the distribution of ε_t is fixed, $b(L)$ completely characterizes the statistical distribution of income. Importantly, though, the agent forecasts the future using only the past history of income since the ε_t are not directly observable.

Agents do not know the true income process. $\hat{b}$ denotes a generic income process.

1.1.2 Consumption under ambiguity

The foundation of the model is a standard model of consumption choice with the addition of ambiguity aversion.

Assumption 2 *Agents choose a consumption policy to optimize*

$$\max_{C^{policy}} \min_{\hat{b} \in B} E\left[\sum_{t=0}^{\infty} -\alpha^{-1}\beta^t \exp\left(-\alpha C_t\right) \mid \hat{b}, W_{-1}\right], \tag{6}$$

where C^{policy} represents a rule for consumption as a function of wealth and the history of observables and B is the (compact) set of models the agent deems plausible. E denotes the expectation operator.

A few features of the specification are notable. First, agents have CARA preferences over consumption for the sake of tractability (yielding lemma 1 below). That rules out wealth effects, but since the paper is not concerned with balanced growth, it is not a major drawback. Online appendix 2 shows that our main results go through similarly in a specification where agents have constant relative risk aversion and uncertainty over returns on wealth instead of income.

Second, the ambiguity aversion is of the form introduced by Gilboa and Schmeidler (1989). Agents choose a consumption policy under the assumption that, whatever policy they set, nature will choose the income process that yields the lowest expected utility. In other words, agents choose a consumption policy meant to be robust to the worst-case scenario for income dynamics. That worst-case scenario is drawn from a set of possible models B, which is typically exogenous in the literature. Unlike Hansen and Sargent (2007), who focus on uncertainty about the distribution of the shocks, ε, our focus here is on how agents learn about dynamics, b, similar to Bidder and Dew-Becker (2016) and Hansen and Sargent (2010, 2017).

Third, note that the ambiguity aversion is not itself dynamic. Agents choose a consumption policy and worst-case model timelessly, optimizing over expected utility similarly to a date-0 problem (though the expectation here is unconditional, not depending on a prior income history). Were the problem fully dynamic, time consistency would be a concern. The analysis will show that uncertainty about low-frequency income dynamics is the primary driver of the model, and those

frequencies are the slowest to learn about, making the assumption that the model $\hat{b} \in B$ is chosen once and for all time not entirely unreasonable.

The minimax theorem implies that the maximization and minimization in (6) can be reversed. Intuitively, (6) represents a zero-sum game with a Nash equilibrium, so that C^{policy} and nature's choice of a $\hat{b}$ are best responses to each other. That means that the equilibrium C^{policy} is optimal for the equilibrium $\hat{b}$. Optimal consumption under CARA preferences has been widely studied and, conditional on a model, the standard results hold here: agents consume the annuity value of human and financial wealth, where human wealth depends on the forecast of future income, and hence $\hat{b}$. The full consumption function is analyzed in section 5.1.

More important to us is nature's minimization problem. Using the solution for optimal consumption yields:

Lemma 1 *Expected utility is a decreasing function of* $\hat{b}\left(R^{-1}\right)^2$*, so that*

$$\arg\min_{\hat{b}\in B}\max_{C^{policy}} E\left[\sum_{t=0}^{\infty} -\alpha^{-1}\beta^t \exp\left(-\alpha C_t\right) \mid \hat{b}, W_{t-1}\right] = \arg\min_{\hat{b}\in B} -\hat{b}\left(R^{-1}\right)^2. \tag{7}$$

where, notationally, $\hat{b}\left(R^{-1}\right) = \sum_{j=0}^{\infty} b_j R^{-j}$***.***

Proof. See appendix A. ■

Since income processes are ranked according to $-\hat{b}\left(R^{-1}\right)^2$, that is the key statistic that drives the analysis throughout the paper and it will appear repeatedly. $\hat{b}\left(R^{-1}\right)$ is the discounted sum of the impulse response function of Y_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.

Lemma 1 represents the baseline ambiguity aversion solution. The worst-case model is the $\hat{b} \in B$ with the highest risk, $\hat{b}\left(R^{-1}\right)^2$. Ambiguity averse agents therefore choose consumption policies that are optimal when income is risky in the sense that shocks to its NPV are large.

A simple way for NPV shocks to be large is for income to have a highly persistent component – i.e. $\hat{b}_j > 0$ for many values of j – which is the basis of the results in Hansen and Sargent (2005, 2017) and Bidder and Dew-Becker (2016) that ambiguity averse agents tend to focus on models with excess persistence. Those papers, however, consistent with the rest of the ambiguity aversion literature, take the set of models, B, as exogenously given. We now endogenize B.

1.2 Information and beliefs

Where this paper contributes to the literature is in allowing the agent to acquire information. We assume agents are able to acquire signals about the true income process, b. The signals are denoted by x, and their precision by τ. The set of plausible models is then written as $B\left(x;\tau\right)$. The signals are not themselves a choice, only their precision. Conditional on τ and b, x is random.

We begin by stating the basic form of the information acquisition problem.

Assumption 3 *Agents choose τ as the solution to*

$$\min_{\tau} E\left[\max_{\hat{b}\in B(x;\tau)} \log \hat{b}\left(R^{-1}\right)^2 \mid \tau\right], \tag{8}$$

subject to a cost of information defined below.

The expectation here is taken over the signals, x. The assumption says that agents choose signal precisions to optimally constrain nature, in the sense of limiting the extent to which nature can choose an unfavorable model. In general, more precise signals will make the set $B(x;\tau)$ smaller and thus reduce (in expectation) the maximum income risk that agents think is plausible. That is their basic motivation for gathering information.

Since the signals are random conditional on τ, there is an element of choice under uncertainty. We show below that assuming τ is chosen to maximize expected $\log \hat{b}\left(R^{-1}\right)^2$ makes the model analytically tractable, and has the advantage of yielding a form of risk-neutrality over the signals. The log transform in (8) affects only choice under uncertainty about the realizations of the signals, x. It is irrelevant to both the optimal consumption and ambiguity aversion, and therefore cannot be determined purely from consumption behavior or knowledge of the set B agents fear. Instead, it determines preferences over the random realizations of the signals. We therefore use a convenient functional form that allows for closed-form solutions.

2 Information acquisition technology

This section describes the structure of the signals x, their precisions, τ, and how they map into the set of plausible models, $B(x;\tau)$.

In general, estimates of dynamics in the time domain, whether in terms of lag polynomials or autocovariances, are correlated across lags. In addition to the facts that these objects are infinite dimensional and have highly nontrivial constraints (positive definiteness for the autocovariances, invertibility for a and b), the complicated correlation structure makes the analysis extremely difficult in the time domain. This section introduces a transformation into the frequency domain that substantially simplifies the analysis, then sets up the information acquisition technology, and finally shows how information is used to construct a set of plausible models.

A standard toolkit has been developed in the literature for studying learning based on independent Gaussian signals. The environment is chosen so that those tools apply directly here.[8] The frequency domain analysis, information acquisition technology, and choice of priors combine to yield a linear-quadratic optimization that has analytic and interpretable solutions. There are other dimensions of learning not considered here, such as filtering of latent states (e.g. Kasa (2006) and Guvenen (2007)).

[8] For example, the final information acquisition problem resembles those studied in Kacperczyk, van Nieuwerburgh, and Veldkamp (2016) and Crouzet, Dew-Becker, and Nathanson (2018).

2.1 The spectrum and income risk

The spectral density of the income process, $\exp f(\omega)$, is the Fourier transform of the autocovariances:

$$\exp f(\omega) \equiv \sum_{j=-\infty}^{\infty} \cos(\omega j) \operatorname{cov}(Y_t, Y_{t-j}). \tag{9}$$

As f is periodic, we may restrict attention to the domain $\omega \in [0, \pi]$. There are one-to-one mappings between f, the autocovariances, and the Wold representation, b, so f fully represents the income process.

The key feature of the spectrum is that it is a variance decomposition for income in terms of fluctuations at different frequencies:

$$var(Y_t) = \frac{1}{\pi} \int_0^{\pi} \exp(f(\omega)) \, d\omega. \tag{10}$$

$\exp(f(\omega))$ measures the contribution of fluctuations in income at frequency ω to the total variance of income. The relative magnitude of f across frequencies determines the extent to which variation in income is driven by low- versus high-frequency fluctuations. An AR(1) process with an autocorrelation near 1 has a spectrum whose mass is isolated at low frequencies, whereas a process that features reversals, such as $Y_t = \varepsilon_t - (1/2)\varepsilon_{t-1}$, has a spectrum with mass concentrated at high frequencies (those near π).

As with b and $\hat{b}$, f is the true log spectral density of income and generic spectra are denoted by $\hat{f}$. There is a surprisingly simple mapping between the spectrum, $\hat{f}$, and the measure of income risk that determines utility, $\hat{b}\left(R^{-1}\right)^2$:

Lemma 2 *For a log spectrum $\hat{f}$ with associated Wold representation $\hat{b}$,*

$$\log \hat{b}\left(R^{-1}\right)^2 = \frac{1}{\pi} \int_0^{\pi} Z(\kappa) \hat{f}(\kappa) \, d\kappa \tag{11}$$

$$\text{where } Z(\kappa) \equiv 1 + 2\sum_{j=1}^{\infty} \cos(\kappa j) R^{-j}. \tag{12}$$

Proof. This is known as the Poisson representation in complex analysis. Insert R^{-1} for z in equation 10.2.10 of Szegő (1939) or equation 2.11 of Inoue and Kasahara (2006). ■

This result shows that utility decreases linearly in $\hat{f}$ – i.e. with the variance of income growth – and the function $Z > 0$ determines the importance of fluctuations at each frequency. The left-hand panel of figure 1 plots Z for an annual calibration with $R = 1.025$. The mass of Z primarily lies on extremely low frequencies, so what matters for the agent's utility is the magnitude of the spectral density at those frequencies.[9]

[9] 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

Lemma 2 drives the rest of the analysis, as it shows that the feature of models that agents worry about most, in the sense that it can damage utility the most, is low-frequency volatility. While the idea that long-run income shocks are most important is a common intuition, following naturally from the permanent income hypothesis, lemma 2 (combined with lemma 1) quantifies *exactly* how utility depends on fluctuations at all frequencies, and it is a result that has not appeared previously in the economics literature.

2.2 Learning about the spectrum

We assume that agents gather information about their income process from a large dataset that reports the income histories of many people, all of whom have the same parameters determining their income processes (i.e. the same f and hence b), but different realizations (different ε's). That dataset can be thought of as representing the information that people can get from talking to family members, teachers, or other mentors who are old enough to have long income histories.

Using standard time series methods (e.g. Brillinger (1981), chapter 5), each income history from that database can be used to provide an estimate of the spectral density of income at one or many frequencies. Specifically, define

$$\tilde{f}_i(\omega) \equiv \log \left| T_i^{-1} \sum_{t=1}^{T_i} Y_{i,t} \exp(i\omega t) \right|^2 + \varrho, \tag{13}$$

where T_i is the length of income history i in the database and ϱ is Euler's constant. $\tilde{f}_i(\omega)$ is an estimator of $f(\omega)$ in that, as $T_i \to \infty$,

$$E\left[\tilde{f}_i(\omega)\right] \rightarrow f(\omega) \tag{14}$$

$$cov\left(\tilde{f}_i(\omega_1) - f(\omega_1), \tilde{f}_i(\omega_2) - f(\omega_2)\right) \rightarrow \frac{\pi^2}{6} 1\{\omega_1 = \omega_2\} \tag{15}$$

$\tilde{f}_i(\omega)$ is an unbiased estimator of f, its errors are uncorrelated across frequencies, and the error variance is independent of both the frequency and the length of the particular history, T_i. While these are asymptotic results, simulations of our benchmark calibration in online appendix 5 show they are highly accurate in small samples.[10] The existence of an estimator with these useful properties is the key reason to perform the analysis in the frequency domain.

To obtain information about the log spectrum of income at some frequency ω, the agent calculates the sample spectrum $\tilde{f}_i(\omega)$ for a number $\tau(\omega)$ of the income histories from the dataset. Treating the average of those sample spectra as approximately Normal (i.e. appealing to the Central Limit theorem) motivates to the following assumption.

loading of utility on frequencies for the level of income is the same as for the first difference.

[10] The result follows from Brillinger (1981) theorem 5.2.6 combined with the continuous mapping theorem. At frequencies 0 and π, the variance doubles. Frequency 0 does not appear in our analysis, and we ignore the doubling at π as it is quantitatively irrelevant, being just a single point.

Assumption 4 *The agent receives signals* $\{x(\omega_j)\}_{j=1,\ldots,n}$ *that are distributed as*

$$x(\omega_j) \sim N\left(f(\omega_j), \tau(\omega_j)^{-1}/d\omega\right) \tag{16}$$

where $\omega_j \equiv \pi j/n$, $d\omega \equiv \pi/n$, *and the errors are uncorrelated across frequencies. The cost of those signals is, for a constant* θ,

$$\theta \sum_{j=1}^{n} \gamma(\omega_j)\tau(\omega_j)\, d\omega. \tag{17}$$

For technical reasons (to avoid infinite information flows, for example), we assume that the agent gains information on the spectrum on the uniform discretization of $[0, \pi]$ given by $\omega_j = \pi j/n$, and we take n as large. We scale the variances by $d\omega$ so that they can be interpreted as the information density at each point, and ignore the $\pi^2/6$ term for simplicity.

The agents also face a cost of gathering information, which comes from the number of income histories that they use at each frequency. Obtaining $\tau(\omega_j)$ observations at frequency ω_j requires calculating $\tau(\omega_j)$ inner products (equation (13)). If τ differs across frequencies, that means that the agent calculates the sample spectrum for more income histories at some frequencies than others. That is, they have many income histories to examine, but for frequencies that they learn less about, they only estimate the spectrum using a small number of them, and the effort saved is allocated elsewhere.

Allowing the cost to vary across frequencies according to $\gamma(\omega)$ makes some frequencies more expensive to learn about than others. In the main results, $\gamma(\omega) = 1$, and we examine the more general case in section 6. The case of a constant γ, in which all frequencies are equally difficult to learn about, is a standard benchmark in the time series literature. The nearly universal result is that estimates of the spectral density have identical variances across frequencies; see Brillinger (1981), Priestley (1981) and Hamilton (1994) for textbook treatments.

An important caveat to the statistical results, though, is that the lowest frequency that can be estimated from a given history depends on the history's length, T_i. A 50-year income history can only directly reveal information about cycles that last 50 years or less. The assumption that an agent can potentially obtain signals at all frequencies means that they can find arbitrarily long income histories in their data. That implication is in fact model consistent since, as is standard, the agents in the model are infinitely lived.

An alternative interpretation of the model is that agents are not infinitely lived, but rather have a constant probability of death in each period (so that effective geometric discounting arises from a combination of the death rate and the rate of pure time preference). Even in that case, some nonzero number of agents live at least T_i periods for any finite T_i. But finding such long-lived agents becomes progressively harder as the required T_i grows. Section 6 enriches the model to account for the idea that finding people with income histories long enough to be informative about the very lowest frequencies should be asymptotically difficult. It also examines a more extreme case where there is a lower bound to the frequencies agents can learn about (equivalently, that there is a $\bar{T}$

such that $T_i \leq \bar{T}$). The benchmark results focus on the $\gamma(\omega) = 1$ case due to its prominence in the literature and the relative ease of its analysis.

A natural baseline in the absence of optimization is for an agent to allocate equal information costs to all frequencies, so that $\tau(\omega_j) \propto \gamma(\omega_j)^{-1}$. When $\gamma(\omega_j) = 1$, this corresponds to obtaining signals with equal variances at all frequencies, which is the textbook time series result. We therefore refer to that as the statistical benchmark allocation.

2.2.1 Relationship with rational inattention

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 it. Then the noise in the signals 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. Such a specification also makes the benchmark with $\gamma(\omega) = 1$ natural since a mental information processing constraint need not bind especially tightly on any particular frequency.

In terms of the literature, the signal structure we analyze is highly similar to that in Kacperczyk, Van Nieuwerburgh, and Veldkamp (2016) in that agents receive signals with normally distributed errors and they are constrained by the total precision of the signals. This constraint is most natural when each independent observation of the spectrum is equally costly to obtain. Sims (2003) proposes an alternative constraint based on information flow or entropy. In our setting, the total entropy of the signals is $\sum_{j=1}^{n} \log(\tau(\omega_j)\, d\omega)$, so high-precision signals are relatively less costly under an entropy constraint. Section 6 provides results for that alternative cost function.

That said, the setting here is more restricted than fully general rational inattention models: the signal errors are Gaussian (motivated by the Central Limit theorem) and uncorrelated across frequencies (motivated by the properties of statistical estimates of the spectrum). In the most general form of the models that Sims (2003) studies, those restrictions need not hold, but they are commonly imposed elsewhere, as in Kacperczyk, Van Nieuwerburgh, and Veldkamp (2016).

2.3 Priors and model plausibility

The previous section referred to ambiguity over models $\hat{b}$ in a set $B(x;\tau)$. Since there is a one-to-one mapping between Wold representations $\hat{b}$ and spectra $\hat{f}$, we can equivalently refer to models $\hat{f}$ in a set $F(x;\tau)$. Given a prior for f and knowing the distribution of the signals x, an agent can use Bayes' rule to calculate the probability that some model $\hat{f}$ is the true model. We assume that the set F is those models whose probability of being the truth is above some cutoff.

Those probabilities cannot be calculated without a prior. A first idea might be to use a flat prior, as that might impose minimal structure. In that case, though, an agent's posterior mode for the spectrum would simply be $\hat{f} = x$. That estimate has the property that it is most variable – and so most complex – exactly where the agent has the least information. Moreover, as the discretiza-

tion becomes small – $d\omega \to 0$ – the estimate has unbounded variation, making it economically implausible.

A more natural situation is for agents to have a prior that enforces some simplicity on the spectrum. We assume that agents believe the log spectrum is likely to be smooth in the sense that its differences across frequencies have limited variation, rather than fluctuating wildly. Following Shiller (1973) and others, the prior is represented by a penalty on variability.[11] The most plausible models have perfectly flat spectra – white noise – while the least plausible have highly variable spectra. Given assumption 4, the log posterior probability of a model $\hat{f}$ is equal to

$$P\left(\hat{f} \mid x, \tau\right) = \underbrace{-\frac{1}{2}\sum_{j=1}^{n}\left(x\left(\omega_j\right) - \hat{f}\left(\omega_j\right)\right)^2 \tau\left(\omega_j\right) d\omega}_{\text{Data log likelihood}} \underbrace{- \frac{\lambda}{2}\sum_{j=2}^{n}\left(\frac{\hat{f}\left(\omega_j\right) - \hat{f}\left(\omega_{j-1}\right)}{d\omega}\right)^2 d\omega}_{\text{Smoothness prior}} + \text{constants}. \tag{18}$$

The parameter λ determines the strength of the prior. When the quality of information, τ, grows, the prior becomes relatively less important and agents focus on more complicated models that track the data more closely. So complexity only arises when people have a wealth of information. The frequencies at which agents have the best signals are also the frequencies at which they will potentially have the most complicated models. At frequencies agents ignore in the sense of selecting small τ, they will tend to use models with spectra close to flat across frequencies.

The posterior probabilities (18) lead to the following assumption on F:

Assumption 5 *The set of "plausible" models is*

$$F\left(x;\tau\right) = \left\{\hat{f} : P\left(\hat{f} \mid x, \tau\right) \geq \bar{p}\right\}. \tag{19}$$

and subject to the condition that $\hat{f}$ is a step function on the discretization $\{\omega_j\}_{j=1}^{n}$

Agents assume that nature can choose models that are not too inconsistent with the data in the sense that their probability of being true conditional on x is above some cutoff. Another way to state the assumption is that, compared to the maximum likelihood estimator of the spectrum, nature can impose an alternative model that cannot be rejected on the basis of a (penalized) likelihood ratio test at some confidence level, depending on $\bar{p}$. The step function simply continues the discretization from above and becomes unimportant as n becomes large.

Hansen and Sargent (2007) study an alternative measure of plausibility, assuming that agents have some exogenously specified benchmark model and that nature is constrained to choose a

[11] The smoothness prior is often explicitly jutified as a belief in simplicity. Shiller (1973), the first application of such a prior, says "[i]n most applications...the researcher will feel that...the lag coefficients should trace out a 'smooth' or 'simple' curve." Akaike (1979) and Kitagawa and Gersch (1985, 1989) use frequency domain priors almost identical to ours. We show below that the smoothness prior also imposes smoothness on the AR and MA coefficients. That white noise is treated as the most plausible is also sensible from an information theoretic perspective since Gaussian white noise has the greatest Shannon entropy among all time series processes with a given variance.

nearby alternative based on the Kullback–Leibler (KL) divergence. The key differences between our set F and that in the Hansen–Sargent framework are that we replace their exogenous benchmark model by the endogenous signals x and that deviations between an alternative model and x are weighted by the chosen precisions τ, whereas the KL divergence puts equal weight on deviations at all frequencies (Dahlhaus (1996)).[12] It is the endogeneity of τ, and hence the set of plausible models, that represents our contribution. In the robust control framework, the set of alternative models is exogenously fixed.

In addition to the smoothness prior, we also assume that agents are able to express a prior mean over possible models. In the absence of any information about the world, they believe the average spectrum is flat at $\bar{f}$. This assumption is introduced so that it is possible for the agent to calculate expectations for $\hat{f}$ prior to observing signals. The belief about the level is associated with infinite variance, though, so it does not appear in the posterior, P. The existence of a prior mean is necessary for calculating the optimal τ, but its level is irrelevant and it has no implications for the model.

3 Solution

All three optimizations in the preferences – the consumption policy, nature's choice of a model, and the information decision – are analytically solvable. There is little work that obtains closed-form solutions for optimal consumption under model uncertainty and rational inattention, and the fact that the model can be solved when model uncertainty and attention are themselves endogenous is even more surprising.

The consumption part of the optimization was already solved in section 1.1.2 with the result that models are ranked according to $\hat{b}\left(R^{-1}\right)^2$ (section 5.1 provides further details). This section solves nature's minimization over models $\hat{f} \in F$ and then finds the agents' optimal choice for τ.

3.1 Nature's minimization

Lemma 2 says that the $f^w \in F$ that maximizes $\int_0^\pi Z(\kappa)\hat{f}(\kappa)\,d\kappa$ also maximizes income risk, and that optimization is a straightforward linear problem. This section reports the solution using vector notation.

Recall that the agent's information and nature's choice of a model are both defined on the discretization $\{\omega_j\}$. The optimization can be studied in a vector form using the values on just those points, so we define a vector (in boldface) $\mathbf{f}^w(x;\boldsymbol{\tau}) \equiv [f^w(\omega_1;x,\tau),...,f^w(\omega_n;x,\tau)]'$ and also $\mathbf{Z}$ taking the same form (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). $diag(\cdot)$

[12] Note also that the KL divergence, which is sometimes also called an entropy distance, is separate from entropy as a measure of information flow to the agents. The KL divergence is used in the robust control setting to determine the set of plausible models, F. The relative entropy used in rational inattention models measures information flows – τ here. Section 6 discusses how the analysis changes under the use of the Shannon entropy to measure information flow.

is an operator that creates a matrix with its argument on the main diagonal and zeros elsewhere. We then have

Proposition 1 *Nature's optimization (7) has the associated Lagrangian*

$$\max_{\hat{f}} \int Z(\omega)\,\hat{f}(\omega)\,d\omega - \psi P\left(\hat{f} \mid x, \tau\right) \tag{20}$$

where ψ is a Lagrange multiplier. The worst-case model that solves (20) is

$$\mathbf{f}^w(x;\boldsymbol{\tau}) = \left(I_n - \lambda diag\left(\boldsymbol{\tau}^{-1}\right) D\right)^{-1}\left(\psi diag\left(\boldsymbol{\tau}^{-1}\right)\mathbf{Z} + x\right) \tag{21}$$

where I_n is the $n \times n$ identity matrix and D is a differencing matrix of the form

$$D \equiv \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & & \vdots \\ 0 & 1 & -2 & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix} d\omega^{-2}. \tag{22}$$

Proof. See appendix B ■

While the notation is somewhat involved, the result here is simple: f^w is a linear function of x and Z. The worst-case spectrum is shifted up – risk is higher – compared to the agent's signals, x, by an amount that decreases with the precision of the signals, increases with the pain associated with the particular frequency, and increases with the degree of ambiguity aversion, represented by ψ. The full solution is reported here because it shows exactly what model the agent uses for decisionmaking, and underlies all of our main results. The vector notation in proposition 1 is relevant for the solution, and it appears in the derivations in the appendix, but it is not used in the remainder of what follows – the analysis involves either individual frequencies, or continuous limits (i.e. $n \to \infty$).

Before analyzing the implications of proposition 1 in detail, we first solve the agent's optimal τ to help frame the effects of information choice on consumption behavior.

3.2 Optimal information choice

Combining assumptions 3 and 4 with lemma 2, the optimization problem for τ becomes

$$\max_{\tau} E\left[\min_{\hat{f}\in F(x;\tau)} -\int_0^{\pi} Z(\omega)\,\hat{f}(\omega)\,d\omega\right] - \theta\sum_{j=1}^{n}\gamma(\omega_j)\,\tau(\omega_j)\,d\omega. \tag{23}$$

Agents choose the signal precisions τ to minimize how bad a model nature can choose for them, where "bad" here is (from lemma 1) measured by long-run risk, $\hat{b}\left(R^{-1}\right)^2$.[13]

Proposition 2 *The optimal information policy that solves (23) when* $\gamma(\omega) = 1$ *is*

$$\tau^*(\omega_j) = \underbrace{\theta^{-1/2}}_{Cost\ of\ info.} \times \underbrace{\psi^{1/2}}_{Ambiguity\ aversion} \times \underbrace{Z(\omega_j)}_{Utility\ weights}. \tag{26}$$

Proof. See appendix C. ■

Agents optimally gather information exactly in proportion to Z, learning the most about the frequencies that are most important for utility. In terms of an adversarial game with nature, the agent chooses precision to constrain nature most at the frequencies that are potentially most painful. The parameters θ and ψ determine the scale of τ^*. When information is less costly or agents are effectively more ambiguity averse – θ falls or ψ rises – agents acquire more more precise signals. To see the implication of proposition 2 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.

The remainder of the paper analyzes the implications of the solution for the types of models that agents optimally use and how those choices affect observable consumption behavior.

4 Behavior of the model agents use

We have two relevant cases for τ. The utility-optimal information policy, $\tau^*(\omega)$, says that it is proportional to $Z(\omega)$, while the statistical benchmark in the absence of optimization is to set $\tau(\omega)$ to equal a constant. We focus on two key results for f^w under those policies:

1. **Optimal learning eliminates excessive extrapolation:** Without an optimal information 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. **Agents make mistakes primarily about the transitory component of income:** Under the optimal policy, agents use models that tend to deviate from the truth more at high than at low frequencies. That behavior does not appear under the non-optimal information policy.

This section derives those results theoretically and examines them in numerical simulations of the model. Section 5 examines how those results map into consumption behavior.

[13] The optimization has been laid out in three separate steps – the consumption/savnigs choice, nature's choice of a model, and the agents' choice of signal precisions, but it can also be written in a single line as

$$\max_{\tau} E\left[G\left(\max_{C^{policy}} \min_{\hat{f}\in F(x;\tau)} E\left[\sum_{t=0}^{\infty} -\alpha^{-1}\beta^t \exp\left(-\alpha C_t\right) \mid \hat{f}\right]\right)\right] \tag{24}$$

$$\text{where } G(x) \equiv -\log\log\left(-\alpha(1-R)(\beta R)^{1-R}\, x\right) \tag{25}$$

4.1 Optimal learning eliminates excessive extrapolation

Taking an expansion around an infinite level of precision, appendix C.1 derives the following first-order approximation in the continuous limit of the problem ($d\omega \to 0$) for arbitrary τ:

$$E\left[f^w\left(\omega; x, \tau\right) - f \mid f\right] \approx \psi\tau\left(\omega\right)^{-1} Z\left(\omega\right) + \lambda\tau\left(\omega\right)^{-1} f''\left(\omega\right). \tag{27}$$

Equation (27) yields our first important result. In the statistical benchmark case where τ is constant across frequencies, f^w is biased in the direction of $Z\left(\omega\right)$. 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 is more persistent than the truth. 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, 2017). Intuitively, since highly persistent models lead to the lowest utility by driving $\hat{b}\left(R^{-1}\right)^2$ up, agents naturally fear them.

Equation (27) also yields the more important part of the result, though, which is that under the optimal policy, τ^*, there is no systematic bias towards either under- or over-extrapolation. Specifically,

$$E\left[f^w\left(\omega; x, \tau^*\right) - f \mid f\right] \approx \psi^{1/2}\theta^{1/2} + \lambda\tau^*\left(\omega\right)^{-1} f''\left(\omega\right). \tag{28}$$

Since $\tau^*\left(\omega\right) \propto Z\left(\omega\right)$, 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, $\psi^{1/2}\theta^{1/2}$.

When f^w deviates from f by only a constant, the two models 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 $E\left[f^w\right]$ is the log spectrum for an AR(1) also with persistence ρ, but with innovations that have a greater variance. On average then, the worst-case Wold representation, $b^w\left(L\right)$, is biased up only in its constant, b_0^w, while all the lag coefficients are unbiased relative to the true model $b\left(\omega\right)$.

Equation (28) 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 entirely canceling out the effect of ambiguity.

This result stands in conflict with recent work that argues that ambiguity aversion and information processing constraints lead to overextrapolation (Fuster, Hebert, and Laibson (2012), Bidder and Dew-Becker (2016), and Hansen and Sargent (2016)). What we find here is that when people are able to choose what aspects of income to learn about, they naturally focus on the low frequencies, since those are most important for utility. But it is precisely that focus that then eliminates any bias towards excessive extrapolation.

4.1.1 Numerical example

To make the results more concrete, we consider a simple numerical example. Suppose income is truly i.i.d. over time, $Y_t = \varepsilon_t$, so that the true model has zero persistence. Since $f''(\omega) = 0$, the second term in equations (27) and (28) is equal to zero. The left-hand panel of figure 2 plots the true (flat) log spectrum $f(\omega)$ along with the mean worst-case spectra under the optimal information policy τ^* and for the statistical benchmark in which τ is constant across frequencies (the calibration is set so that they have equal total precision: $\sum_j \tau^*(\omega_j) = \sum_j \tau$), which we denote with $\bar{f}^w_*$ and $\bar{f}^w_F$, respectively. The figure shows that $\bar{f}^w_*$ is shifted up by a constant compared to f, while $\bar{f}^w_F$ actually has a significantly different shape, with a peak at low frequencies indicating persistence in income.

The right-hand panel of figure 2 plots the impulse response functions (the b's) associated with the three models. Since income is truly i.i.d., $b_j = 0$ for $j \geq 1$ under the true model. Under the optimal information policy with model uncertainty, the only thing that changes on average is that b_0 becomes larger – people fear a higher variance, but they do not on average act as though income actually has any persistence. Under the statistical benchmark, though, there is clearly persistence in income: the impulse response is consistently positive after the initial impact. Figure 2 thus illustrates our first basic result. While ambiguity aversion and model uncertainty can often drive agents to act as though income is excessively persistent, that result is delicate: it disappears when people can allocate attention and information acquisition optimally.

4.2 Agents make mistakes about the transitory component of income

The primary mistakes in the agent's worst-case model come from the term in (27) involving $f''(\omega)$. That part of the formula is driven by the agent's smoothness prior. In the face of noisy data, agents estimate the spectrum of income by smoothing information across frequencies. Since $f^w(\omega; x, \tau)$ 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.

In that sense, the agents also have a bias towards simplicity: they use models with smaller variations across frequencies when they have less information.[14] When the true spectrum is in fact complex, in the sense that it has local peaks and troughs, any estimated model will tend to make mistakes in smoothing those peaks out. So the errors appear exactly where f'' is large. The information policy matters here, though, because it determines the frequencies at which those mistakes are concentrated.

Since the optimal information policy gives the agents noisier signals about the spectrum at high frequencies, that is also where they make the largest smoothing errors. In (28), $f''(\omega)$ is multiplied by $\tau^*(\omega)^{-1}$. So when precision is high, the term is scaled down and the worst-case spectrum tracks

[14] That intuition can be formalized. It is possible to show that correlations in the estimated spectrum, $f^w(\omega; x, \tau)$, are higher across frequencies, implying that complexity is lower, in regions where τ is smaller.

the true spectrum closely. But when τ^* is small – at high frequencies – agents do more smoothing across frequencies and make larger mistakes.

4.2.1 Numerical example

To illustrate the errors caused by smoothing, we now consider a richer and more realistic numerical example with multiple peaks in the spectrum. The left-hand and middle panels of figure 3 plot the log spectrum of the data-generating process for income, while the right-hand panel plots the impulse response of income to a shock. The calibration is chosen to have both high-and low-frequency components (see evidence discussed in Kaplan and Violante (2010)). The high-frequency piece – which generates the middle peak in the spectrum – is driven by the fact that a component of the shocks to income reverts: when income rises higher by \$1 today, it is lower on average by 50 cents over the next three periods. That behavior can be caused by forces that shift income over time but have little effect on total lifetime income. For example, many people overpay taxes during the year and then receive refunds (e.g. Souleles (1999)). The low-frequency component of income – the left-hand peak in the spectrum – comes from the fact that the impulse response is persistently positive in the later periods following a shock. This represents a persistent component in income growth, and could come from variation over time in the average growth rate of the economy or the performance of one's employer. The preference parameters are chosen to illustrate the main mechanisms in the model. θ is chosen so that agents make quantitatively large consumption mistakes (see below).[15]

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^*:

$$\tau^F(\omega_j) = \tau^F \equiv n^{-1} \sum_{i=1}^{n} \tau^*(\omega_j). \tag{29}$$

As in the previous example, the choice of the mean for τ^F implies that it has the exact same information cost as τ^*. 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^*.

Figure 3 plots $\bar{f}_*^w$ and $\bar{f}_F^w$ for the two-peak calibration. The average model under the optimal policy, $\bar{f}_*^w$, matches f very well at the lowest frequencies, but it does a poor job of matching the middle-frequency peak in f and also deviates substantially at higher frequencies. The average model under the statistical benchmark, $\bar{f}_F^w$, has the opposite behavior: it matches the middle-frequency

[15] Technically, the impulse response function for income is equal to [1,-0.15,-0.3,-0.15,0,...] plus $0.095\exp(-0.1j)$. It is then scaled so that the standard deviation of consumption growth is 1.56 percent (when initial consumption is equal to 1).

As discussed above, n is intended to be taken as large – it is only used to avoid infinities – so we set it to 4000. $\beta = 0.975$ to represent an annual calibration, and $R = \beta^{-1}$ for simplicity. $\bar{\tau}$, 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. $\psi = 10^{-4}$; $\lambda = 0.00075$; $\bar{\tau} = 405.83$; $\theta = 49.35$. The parameterization is meant to illustrate the qualitative behavior of the model rather than match specific quantitative data. The degree of ambiguity aversion, ψ, has minimal effects on behavior under the optimal information policy, but it matters much more under the flat information policy.

peak and high-frequency behavior well, and in fact matches f well at almost all frequencies, but it fits relatively poorly at low frequencies. That is exactly what the formulas predict: optimal learning causes models to be relatively more accurate at low than high frequencies. Overall, though, $\bar{f}_F^w$ has a much better fit than $\bar{f}_*^w$, with a root mean squared error that is 42 percent smaller, due to the fact that $\bar{f}_F^w$ spreads information evenly across frequencies.

The right-hand panel of figure 3 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. $\bar{b}_*^w$ fails to match the short-run mean-reversion in the income process, while the lag polynomial for the suboptimal information policy, $\bar{b}_F^w$, does not, as predicted by the analytic results. The figure shows that the greater smoothness of $\bar{f}_*^w$ also translates into smoothness in the associated lag polynomial, and in particular errors in the transitory behavior of income. But the figure shows that the optimal policy performs better at longer lags, giving a closer fit to the persistent component of the impulse response function. Since it is the long-run part that determines human wealth, and hence optimal consumption, it is optimal from an expected utility perspective for agents to use models that fit the persistent component at the cost of missing the transitory dynamics.

5 Implications for observable consumption behavior

We now explore the implications of the results in the previous section for the observable behavior of consumption.

5.1 Consumption function

Online appendix 1 shows that given a worst-case model b^w, consumption growth follows

$$\Delta C_t = \left(1-R^{-1}\right) b^w\left(R^{-1}\right)\varepsilon_{t+1}^w + \frac{\alpha}{2}\left(1-R^{-1}\right)^2 b^w\left(R^{-1}\right)^2 + \alpha^{-1}\log\beta R \tag{30}$$

$$\text{where } \varepsilon_{t+1}^w \equiv b^w(L)^{-1} Y_t. \tag{31}$$

Δ is the first-difference operator and $b^w(L)$ is the Wold representation associated with the worst-case model f^w. In the case where agents use the true model, so that $b^w = b$ (i.e. under complete information), 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, $\left(1-R^{-1}\right) b\left(R^{-1}\right)\varepsilon_{t+1}$. When the agent uses a model that differs from the truth, though, ε_{t+1}^w is no longer an i.i.d. process and consumption growth is no longer uncorrelated over time. That is, the agent's estimated shocks, ε^w, are in general serially correlated, which leads to serial correlation in consumption growth, which is suboptimal from a full-information perspective.

The log spectrum of consumption growth is

$$f_{\Delta C}^w(\omega;x,\tau) = \log\left(\left(1-R^{-1}\right)^2 b^w\left(R^{-1};x,\tau\right)^2\right) + f(\omega) - f^w(\omega;x,\tau). \tag{32}$$

When the agent knows the true model, $f^w_{\Delta C}$ is perfectly flat, and we have the usual result that consumption growth is uncorrelated over time and the level of 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 $f(\omega) - f^w(\omega; x, \tau)$. 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.

Using (32), we can immediately map the results in the previous subsections into the spectrum of consumption growth. Specifically, for general information policies and for the optimal policy, plugging (27) into (32) yields

$$\begin{aligned} E\left[f^w_{\Delta C}(\omega; x, \tau) \mid f\right] &\approx E\log\left(\left(1 - R^{-1}\right)^2 b^w\left(R^{-1}; x, \tau\right)^2\right) \\ &\quad -\psi\tau(\omega)^{-1} Z(\omega) - \lambda\tau(\omega)^{-1} f''(\omega), \qquad (33) \\ E\left[f^w_{\Delta C}(\omega; x, \tau^*) \mid f\right] &\approx E\log\left(\left(1 - R^{-1}\right)^2 b^w\left(R^{-1}; x, \tau^*\right)^2\right) \\ &\quad -\psi^{1/2}\theta^{1/2} - \lambda\tau^*(\omega)^{-1} f''(\omega). \qquad (34) \end{aligned}$$

Again, the information policies differ in two key ways. First, comparing the terms $\psi\tau(\omega)^{-1} Z(\omega)$ and $\psi^{1/2}\theta^{1/2}$, there are no systematic deviations of consumption growth from white noise under the optimal information policy. Under other policies, though, since people overextrapolate income shocks, consumption is actually mean reverting in the long-run – there is a trough in $f^w_{\Delta C}$ at frequency zero. Intuitively, overextrapolation causes people to consume more than they can afford (more than human wealth) following positive shocks. Eventually, then, they must reduce consumption, causing long-run mean reversion. So the observable prediction of the model is that we actually *should not* observe long-run mean reversion in consumption growth. By the same token, people should also not underreact to shocks (as under rational inattention), which would lead to long-run persistence in consumption growth.

The second class of mistakes is the smoothing errors due to the term $\lambda\tau(\omega)^{-1} f''(\omega)$. This term says essentially that variation in the spectrum of income that the agent is not aware of passes directly into consumption growth. When f'' is negative, for example, there is a local peak in the spectrum of income, and the spectrum of consumption growth then is also relatively high. Again, these errors are scaled by the precision of signals. The model predicts that consumption should track income relatively more closely – have a similar impulse-response function – at high than low frequencies. Transitory variation in income, such as the shifts in income over time studied by Souleles (1999), is predicted to pass directly into consumption. We illustrate that behavior below in a numerical example.

Compared to the behavior under the standard setup with no model uncertainty, our model generates, through limited information, excess sensitivity of consumption to high-frequency shocks to income. This result is not obtained, though, by appealing to some sort of irrationality; rather, it

arises simply from people optimally choosing to focus their attention on low frequencies. Endogenous attention leads to our second difference from the literature, which is that unlike other recent work on model uncertainty (Fuster, Hebert, and Laibson (2012), Bidder and Dew-Becker (2016), and Hansen and Sargent (2016)), the model does *not* predict excessive extrapolation of shocks. The model predicts excess sensitivity to transitory variation in income, but in fact the *correct* sensitivity to the permanent component.

The model also has rather different predictions from rational inattention over state variables (as opposed to rational inattention over model specifications), which suggests that they could be tested against each other empirically. As discussed by Sims (2003), the most prominent prediction of rational inattention is delayed reaction to shocks, due to the fact that people observe the shocks imperfectly. If income rises permanently, Sims (2003) shows that in general people will take a number of periods to fully realize that such a shock has occurred, meaning that consumption responds slowly to permanent shocks to income. Here, on the other hand, agents respond rapidly to permanent shocks because it is precisely the low-frequency part of income that they understand best.

Sims (2003) and Luo (2008) show that rational inattention can also generate excess sensitivity of consumption to income shocks, but the effects are calibration-specific and may be quantitatively small (e.g. the simulations in Sims (2003)). Intuitively, excess sensitivity arises because agents are not able to distinguish permanent from transitory shocks. So to obtain high-frequency mistakes, the rational inattention model must also predict low-frequency mistakes. In our model, though, the prediction of optimal information acquisition is in fact that the same attention choice both induces high-frequency mistakes and eliminates low-frequency mistakes. Furthermore, we see in the next section that the high-frequency mistakes can be quantitatively large and realistic.[16]

5.2 Numerical example

We examine the behavior of consumption under the numerical simulation when income has both transitory and persistent components. Figure 4 plots the log spectra of consumption growth under the various models. $\bar{f}_*^w$ provides a closer fit to the utility optimal consumption spectrum at all frequencies. On the other hand, the statistical information policy produces a spectrum that is flatter – and closer to white noise – across most frequencies, but it has a very large peak at the lowest frequencies. The key question, then, will be which type of deviation – low- or high-frequency – is more relevant for utility.

To see how the fitting errors affect the behavior of consumption growth in the time domain, the right-hand panel of figure 4 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

[16] It is also worth noting that the models in Sims (2003) and Luo (2008) can only be solved under quadratic utility, whereas we are able to accommodate CARA preferences here.

holds, and 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 some of the short-run mean-reversion in income, rising and falling in the first few periods. 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 80 periods of the IRF plotted here, eventually overshooting. So the τ^F policy, counter to what is observed 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.

Those results may also be observed in more standard time series regressions for consumption growth. Table 1 below reports the coefficients from simulated regressions of consumption growth on the predictable and unpredictable components of income growth under the two information policies and also under the full-information optimum.[17]

Information policy	Predictable income	Unpredictable income
τ^*	0.50	1.13
τ^F	0.12	0.72
Full-info. optimum	0	1.11

Table 1. Coefficients from regressions of consumption growth on income growth

The coefficient from the regression of consumption growth on the predictable part of income is of the same order of magnitude as the coefficient on the unpredictable part under τ^*. The model can thus replicate the empirical result that consumption responds strongly to predictable income changes. That value is consistent (by calibration) with the results of Parker (1999) and Souleles (1999), who both find that consumption rises by 0.5 percent following a 1-percent anticipated increase in income.

Under the statistical benchmark, τ^F, on the other hand, that relationship is much weaker, with the response to predictable income being, at 0.12, smaller by a factor of four. It is precisely the fact that agents optimally (under τ^*) fail to learn about high-frequency features of the model that causes them to overreact to predictable parts of income. Furthermore, note that the response of

[17] Here we use the version of the model in which income is difference-stationary. That is, the calibration of the impulse responses and the spectrum are applied to income growth instead of its level. As discussed in footnote 9, the theoretical results go through identically in that case. The difference is simply that then consumption and income have volatilities that are of the same order of magnitude, as observed in the data. The calibration is otherwise identical to what is discussed in footnote 15.

consumption to true income shocks is far closer to the full-information optimum under τ^* than under τ^F. This again demonstrates that τ^* helps agents get long-run responses right.

An alternative way to examine the behavior of consumption in the time domain is to study its autocorrelations. The left-hand panel of figure 5 plots the autocorrelations of consumption growth under τ^* and τ^F. Obviously under the full-information optimum, the autocorrelations are zero. At short lags, the autocorrelations are higher under τ^*. Subsequently, though, the autocorrelations are substantially lower – by nearly a factor of 10. The right-hand panel plots the first autocorrelation of consumption growth over different spans. For a horizon denoted by n on the x-axis, we plot $corr\left(\sum_{j=0}^{n-1}\Delta C_{t+j}, \sum_{j=0}^{n-1}\Delta C_{t-n+j}\right)$. So the figure represents how consumption growth is correlated over neighboring intervals of length n. Consistent with the left-hand panel, for short intervals the correlations are higher under τ^* than τ^F. As we claimed above, though, the figure shows that consumption growth over long periods is substantially less autocorrelated under τ^* than τ^F.

To summarize, this example confirms the analytic results above that the optimal information policy generates 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. It also shows that the model can generate the empirical result that consumption responds to predictable variation in income.

5.3 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, $f^w(\omega; x, \tau^*)$, 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 $f^w_{\Delta C}(\omega; x, \tau^*)$ 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). The shocks studied in those papers essentially shift income over time, exactly as in our numerical example. The results above show that consumption in the model does in fact respond to such variation in income, and that it tracks predictable income variation strongly.

Kaplan and Violante (2014) 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. less than 1 percent 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 mistakes should appear in response to short-duration shocks in our setting, and the empirical research finding violations of the permanent income hypothesis typically studies transitory income shocks.

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.

That implication requires aggregation, though, which is a nontrivial step. Since the consumption function in our model is linear, it will have desirable aggregation properties, but the exact details will depend on how income is driven by aggregate and idiosyncratic shocks at each frequency. Aggregate empirical results are thus not an ideal test of the model. The most direct test would be to measure the extent to which individual consumption growth is close to white noise over long horizons.

An alternative way to test the model, instead of examining consumption, 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. That said, this paper faces the same problem as others in the information acquisition literature that there is no direct data on the type or quantity of information that people have (see Angeletos, Collard, and Dellas (2017) for a related discussion).

6 Alternative information cost specifications

The baseline case has equal information costs across frequencies – $\gamma(\omega) = 1$ – consistent with textbook time series analysis results. There is a common intuition, though, that information about low frequencies should be more costly to obtain. That intuition does not have a formalization, so this section introduces one. We then examine a case where agents are constrained in the total entropy of their signals.

The online appendix reports two extensions of the results here. First, section 3 develops an alternative formalization of the idea that low frequencies are harder to learn about than higher frequencies. In that specification, even though low frequencies are harder to learn about, the information policy remains $\tau^* \propto Z$, as in the baseline. In what follows with $\gamma(\omega) \neq 1$, on the other hand, that will not be true. Second, section 4 examines a case in which there is a $\bar{\omega}$ such that agents cannot directly learn about any frequency $\omega \leq \bar{\omega}$. The results are highly similar to what is reported in section 6.1.

6.1 Costs varying by frequency

Recall that the model of information acquisition is that agents have a database of income histories that they can query. If the people whose incomes are in the database (i.e. the friends, mentors, etc. who the agent learns from) exit the labor force at a constant rate over time, then the length of the histories in the database is naturally geometrically distributed. Specifically, if people exit the labor force in each period with probability δ, then the fraction of people in the dataset who worked for at least k periods is $(1-\delta)^{k-1}$.

In order to get information about cycles that least k periods, the agent needs to look at an income history that is at least k periods long. So to learn about a frequency ω, the agent must find a history lasting $2\pi/\omega$ periods. On average, that will require looking at $(1-\delta)^{1-2\pi/\omega}$ histories. For lower ω, there are fewer sufficiently long histories, and thus less available information. We therefore set in this section

$$\gamma(\omega) = (1-\delta)^{1-2\pi/\omega}. \tag{35}$$

The benchmark results so far correspond to the case where $\delta = 0$. We now study how the results are affected by assuming $\delta > 0$. The only direct effect this has on the model is to change the optimal information policy. The general expressions obtained above for f^w ((21) and (27)) and consumption growth ((30), and (32)) as functions of τ continue to hold. As in the baseline case, we calibrate the parameter θ in this section so that the response of consumption to a one-percent increase in the predictable consumption is 0.5 percent.

6.1.1 Optimal information policy

While the model does not appear to have a closed-form solution in general when $\delta > 0$, there is a solution in the case with no smoothness prior ($\lambda = 0$):

Proposition 3 *The optimal information policy for arbitrary γ and for $\lambda = 0$ is*

$$\tau^{\gamma}(\omega_j) = \underbrace{\gamma(\omega)^{-1/2}}_{1/\textit{Frequency-specific cost}} \times \underbrace{\theta^{-1/2}}_{\textit{Cost of info.}} \times \underbrace{\psi^{1/2}}_{\textit{Ambiguity aversion}} \times \underbrace{Z(\omega_j)}_{\textit{Utility weights}}. \tag{36}$$

Proof. See appendix C.2. ■

The only difference between this result and the main specification is that τ^{γ} now is decreasing in the frequency-specific information costs – agents obtain less information about expensive than about inexpensive frequencies. At the same time, though, they continue to obtain information in proportion to the utility weights. What this result shows, then, is that even with frequency-specific information costs, agents always undo the effects of nature's minimization by setting τ larger where Z is larger, all else equal. At the same time, though, τ^{γ} is obviously smaller when information is more costly, which can cause nature to distort the model at the costly frequencies.

We calibrate $\delta = 0.975$, corresponding to an exit probability of 2 percent, which would imply people have on average a 50-year working life if each period is a year. The top panels of figure 6 then plot the optimal information policies $\tau^* \propto Z$ and

$$\tau^{\gamma}(\omega) = \theta^{-1/2}\psi^{1/2}(1-\delta)^{1-2\pi/\omega} Z(\omega). \tag{37}$$

Both lines again peak at low frequencies, but whereas τ^* peaks at frequency zero, τ^{δ} peaks at a slightly interior frequency (the lines are normalized to have equal information costs). That peak comes at a frequency corresponding to cycles lasting approximately 160 years, though. So while the function γ in this case causes agents to learn less about the very lowest frequencies, the function Z is sufficiently strong that it still causes people to focus their attention on extremely low frequencies.

6.1.2 Behavior of the model agents use

To find the average behavior of the worst-case model with arbitrary γ, we insert the analytic solution for τ^{γ} from the case where $\lambda = 0$ – i.e. ignoring the effects of smoothing on the optimal τ – into the general formula for the bias, (27), to obtain

$$\begin{aligned} E[f^w(\omega; x, \tau^{\gamma}) - f \mid f] \;\approx\; & (1-\delta)^{(1-2\pi/\omega)/2}\theta^{1/2}\psi^{1/2} \\ & +\lambda\theta^{1/2}\psi^{-1/2}(1-\delta)^{(1-2\pi/\omega)/2} Z(\omega) f''(\omega). \end{aligned} \tag{38}$$

The first term again represents the average bias. When the cost of information at low frequencies is larger, the model tends to be biased upward at low frequencies, inducing excessive extrapolation. So we now obtain excess extrapolation, but through a different mechanism than in past work on ambiguity in dynamic models (Hansen and Sargent (2010, 2016) and Bidder and Dew-Becker (2016)). Whereas excess extrapolation in those models appears because low-frequency shocks are most painful to agents, in (38) it appears because low frequencies are most difficult to learn about.

As above, we compare the results under τ^* and τ^{γ} to the case where an equal fraction of

the total information budget is allocated to each frequency. In this case, then, that is $\tau^{F\gamma} \propto (1-\delta)^{-(1-2\pi/\omega)/2}$. The middle panels of figure 6 plot the worst-case spectra under various τ policies now also including τ^{γ} and $\tau^{F\gamma}$.[18] That policy leads to results between the benchmark τ^* and the equal-cost $\left(\tau^{F\gamma}\right)$ policy. At the very lowest frequencies, the τ^{γ} model does not match the true spectrum as well as τ^*, but it still does much better than $\tau^{F\gamma}$. At the middle frequency peak and at higher frequencies, on the other hand, the policy τ^{γ} does a better job of matching the log spectrum than τ^* but still worse than $\tau^{F\gamma}$.

6.1.3 The response of consumption to shocks

The bottom two panels of figure 6 plot the behavior of consumption under the various models. The left-hand panel plots the spectra. The spectral density of consumption growth under the τ^{γ} policy again lies between those of the τ^* policy and the statistical benchmark. While it has a hump at middle frequencies, it is somewhat smaller than that for the optimal policy. The bottom-right panel plots the impulse responses of consumption. The τ^{γ} policy is nearly as effective as the τ^* policy at replicating the optimal initial response of consumption, and it shares some of the excess sensitivity to the short-run variation in income.

The table below calculates the coefficients from regressions of consumption growth on the predictable and unpredictable parts of income, as we did for the benchmark results.

Information policy	Predictable income	Unpredictable income
τ^{γ}	0.50	1.00
$\tau^{F\gamma}$	0.29	0.75
Full-info. optimum	0	1.11

Table 2. Coefficients from regressions of consumption growth on income growth with frequency-specific information costs

The optimal information policy again generates excess sensitivity to income shocks. In this case, compared to the baseline, the response of consumption to unpredictable income shocks is slightly smaller, but it is still economically very close to the full-information optimum. The equal-cost information policy, $\tau^{F\gamma}$, as with τ^{F} above, again generates much smaller responses to both predictable and unpredictable shocks to income.

Overall, when low frequencies are more costly to learn about, the main results are weakened, but by a quantitatively small amount. Agents continue to 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. Agents under both the baseline and the case with frequency-specific information costs make relatively larger mistakes in their models at middle than low frequencies,

[18] For non-zero λ with γ varying across frequencies, $\tau^{\gamma}(\omega) \propto Z(\omega)\gamma(\omega)^{-1/2}$ 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 6 is set up so that the total precision under τ^* is the same as that under τ^{γ} – they differ only in how that precision is allocated across frequencies.

they respond suboptimally strongly to predictable variation income, and they respond nearly optimally to the unpredictable component. The model therefore generates similar overall behavior in beliefs and consumption under both information cost specifications.

6.2 Entropy cost for information

In terms of entropy, the total information flow to the agents in the model can be measured based on the difference between the prior variance that they have at each frequency and the posterior. In our case, since we do not write down a fully specified prior, the information flow cannot be calculated exactly. One way to interpret our agents' prior information, though, would be as a limit in which the prior variance becomes infinite. In such a case, the frequency-by-frequency information flow approaches $\sum_j \log\left(\tau\left(\omega_j\right)\right) d\omega$.

As with the $\delta > 0$ case above, there is not a closed-form solution in the case with entropy costs. However, assuming $\lambda = 0$ again (and with $\gamma = 1$), appendix C.2 shows that the optimal information policy takes the form

$$\tau^{entropy}\left(\omega_j\right) = \theta^{-1}\psi Z\left(\omega_j\right)^2 \tag{39}$$

So whereas in the benchmark we had $\tau^* \propto Z$, with the entropy constraint we find $\tau^{entropy} \propto Z^2$, thus making the focus on the lowest frequencies even stronger and further emphasizing the main results. Intuitively, this result appears because the entropy cost is logarithmic in the precision of the signals, rather than linear, making particularly precise signals relatively less costly in this case than in the benchmark.

7 Conclusion

The basic innovation of this paper is to endogenize the set of models that agents consider in a model of ambiguity aversion. Typically, there is an exogenously specified set of models and agents make decisions as though they will face an unfavorable model. When agents can gather information so as to narrow the set of models, they naturally choose to learn about the features of the world that are potentially most painful. Optimal learning therefore can therefore eliminate some of the most severe consequences of ambiguity aversion.

This paper studies ambiguity and learning in a standard consumption/savings problem. In that setting, it is low-frequency shocks that are most painful. Without learning, agents therefore tend to focus on models with excess persistence, causing them to overextrapolate income shocks. In our model, though, when it is precisely low frequencies that agents choose to learn about. In the case where information is equally costly at all frequencies, they do so in a way that turns out to exactly cancel out the excess extrapolation. More generally, they always focus on the lowest frequencies in such a way as to at least substantially reduce the overextrapolation caused by ambiguity.

An interesting side-effect of the optimal learning is that agents then end up making mistakes at high frequencies. By focusing their attention on low frequencies, they are relatively uninformed

about the transitory variation in income. So optimal information acquisition actually leads agents to consume suboptimally in response to transitory income shocks. We show that the model can generate quantitatively realistic overreactions to predictable transitory variation in income.

In the end, then, the paper contributes to the ambiguity literature by showing that making the set of models over which agents are ambiguous endogenous can have important quantitative and quantitative implications, and it contributes to the consumption literature by providing a fully solvable framework in which agents learn about income dynamics and endogenously make the types of consumption mistakes that have been observed empirically.

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.

Angeletos, George-Marios, Fabrice Collard, and Harris Dellas, "Quantifying Confidence," 2017. Working paper.

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.

Bhandari, Anmol, Jaroslav Borovicka, and Paul Ho, "Survey data and subjective beliefs in business cycle models," 2017. Working paper.

Bianchi, Francesco, Cosmin Ilut, and Martin Schneider, "Uncertainty shocks, asset supply and pricing over the business cycle," 2017. Working paper.

Bidder, Rhys and Ian Dew-Becker, "Long-Run Risk is the Worst-Case Scenario," *The American Economic Review*, September 2016, *106* (9), 2494–2527.

Bidder, Rhys M and Matthew E Smith, "Robust animal spirits," *Journal of Monetary Economics*, 2012, *59* (8), 738–750.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Diagnostic Expectations and Credit Cycles," *Working Paper*, 2016.

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

Brockwell, Peter J. and Richard A. Davis, *Time Series: Theory and Methods*, Springer, 1991.

Caballero, Ricardo J., "Consumption puzzles and precautionary savings," *Journal of Monetary Economics*, 1990, *25* (1), 113–136.

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.

___ **and Argia Sbordone**, "Multivariate Estimates of the Permanent Components of GNP and Stock Prices," *Journal of Economic Dynamics and Control*, 1988, *12(2–3)*, 255–296.

Collin-Dufresne, Pierre, Michael Johannes, and Lars A Lochstoer, "Parameter Learning in General Equilibrium: The Asset Pricing Implications," *The American Economic Review*, 2016, *106* (3), 664–698.

Crouzet, Nicolas, Ian Dew-Becker, and Charles G. Nathanson, "On the effects of restricting short-term investment," 2018. Working paper.

Dahlhaus, R., "On the Kullback-Leibler information divergence of locally stationary processes," *Stochastic Processes and their Applications*, 1996, *62* (1), 139–168.

Drechsler, Itamar, "Uncertainty, Time-Varying Fear, and Asset Prices," *The Journal of Finance*, 2013, *68* (5), 1843–1889.

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.

Epstein, Larry G. and Shaolin Ji, "Optimal Learning and Ellsberg's Urns," 2017. Working paper.

Friedman, Milton, *A Theory of the Consumption Function*, Princeton University Press, 1957.

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

Gabaix, Xavier, "Behavioral Macroeconomics Via Sparse Dynamic Programming." Working paper.

Gersch, Will and Genshiro Kitagawa, "Smoothness Priors Transfer Function Estimation," *Automatica*, 1989, *25* (4), 603–608.

Gilboa, Itzhak and David Schmeidler, "Maxmin Expected Utility with Non-unique Prior," *Journal Of Mathematical Economics*, 1989, *18* (2), 141–153.

Hall, Robert E., "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence," *Journal of Political Economy*, 1988, *86* (6), 971–987.

Hamilton, James Douglas, *Time series analysis*, Princeton university press Princeton, 1994.

Hansen, Lars Peter and Thomas J. Sargent, "Fragile Beliefs and the Price of Uncertainty," *Quantitative Economics*, 2010, *1(1)*, 129–162.

___ and ___ , "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.

Ilut, Cosmin L and Martin Schneider, "Ambiguous business cycles," *The American Economic Review*, 2014, *104* (8), 2368–2399.

Inoue, Akihiko and Yukio Kasahara, "Explicit representation of finite predictor coefficients and its applications," *The Annals of Statistics*, 2006, *34* (2), 973–993.

Jappelli, Tullio and Luigi Pistaferri, "The Consumption Response to Income Changes," *Annual Review of Economics*, 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, "A Rational Theory of Mutual Funds' Attention Allocation," *Econometrica*, 2016, *84* (2), 571–626.

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

Kasa, Kenneth, "Robustness and information processing," *Review of Economic Dynamics*, 2006, *9* (1), 1–33.

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

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

Luo, Yulei, "Consumption Dynamics under Information Processing Constraints," *Review of Economic Dynamics*, 2008, *11* (2), 366–385.

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

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.

Shiller, Robert J., "A Distributed Lag Estimator Derived from Smoothness Priors," *Econometrica*, 1973, pp. 775–788.

Sims, Christopher A, "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.

Szegő, Gabor, *Orthogonal polynomials*, Vol. 23, American Mathematical Soc., 1939.

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," *Review of Economic Studies*, 2006, *73* (3), 823–845.

___ , *Information Choice in Macroeconomics and Finance*, Princeton University Press, 2011.

Wang, Neng, "Precautionary Saving and Partially Observed Income," *Journal of Monetary Economics*, 2004, *51* (8), 1645–1681.

___ , "Optimal consumption and asset allocation with unknown income growth," *Journal of Monetary Economics*, 2009, *56* (4), 524–534.

A Proof of lemma 1

From online appendix 1, the optimal consumption rule is

$$C_t = (R-1) W_{t-1} + -\frac{\alpha}{2} R^{-1} \left(1 - R^{-1}\right) \hat{b} \left(R^{-1}\right)^2 + z(L) \hat{\varepsilon}_t - \alpha^{-1} \frac{\log \beta R}{R-1} \tag{40}$$

for a lag polynomial $z(L)$. The Euler equation immediately implies that

$$E_t \left[-\alpha^{-1} \sum_{j=0}^{\infty} \beta^j \exp\left(-\alpha C_{t+j}\right) \right] = \frac{-\alpha^{-1}}{1-R} \exp\left(-\alpha C_t\right) \tag{41}$$

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

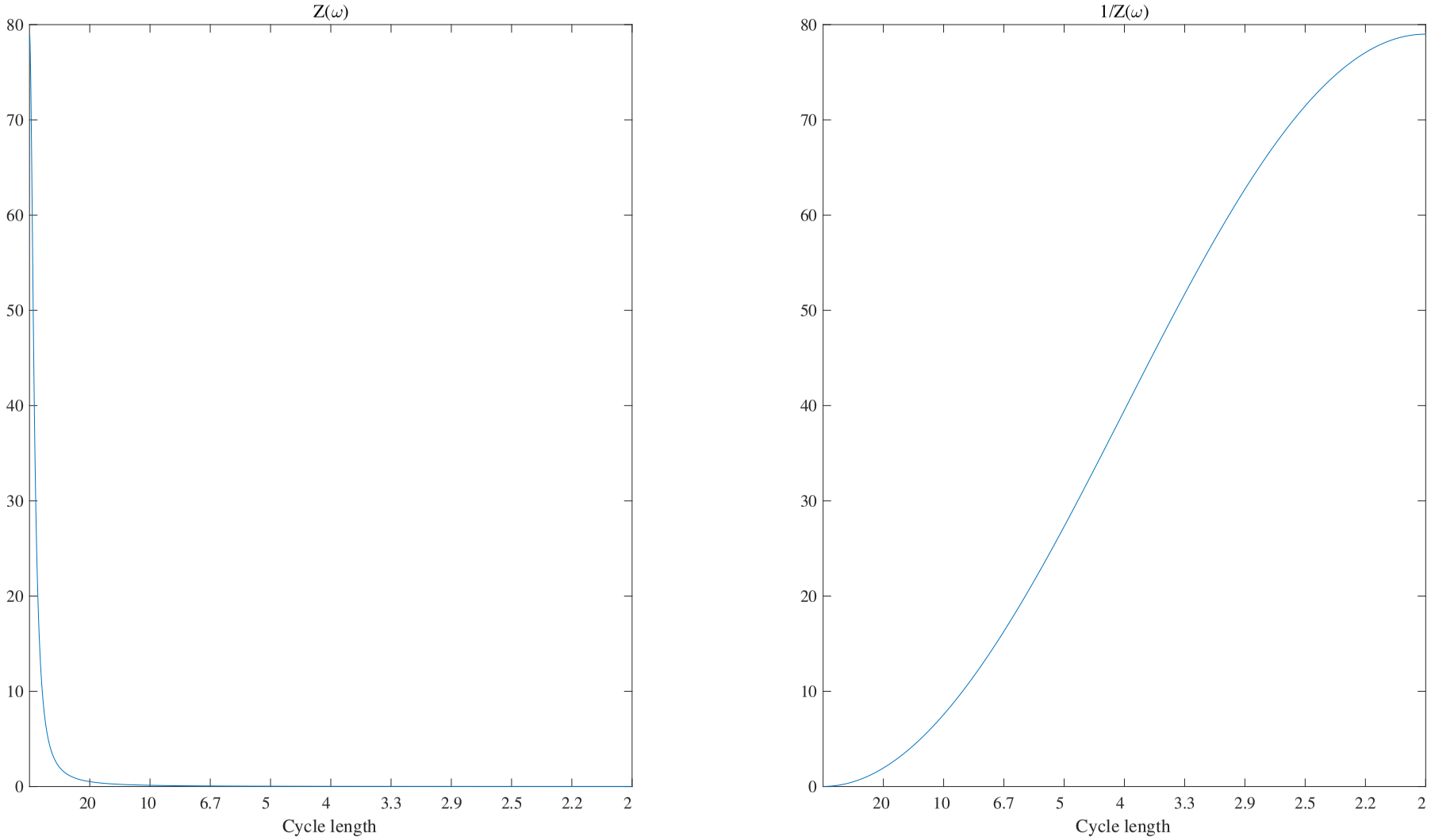
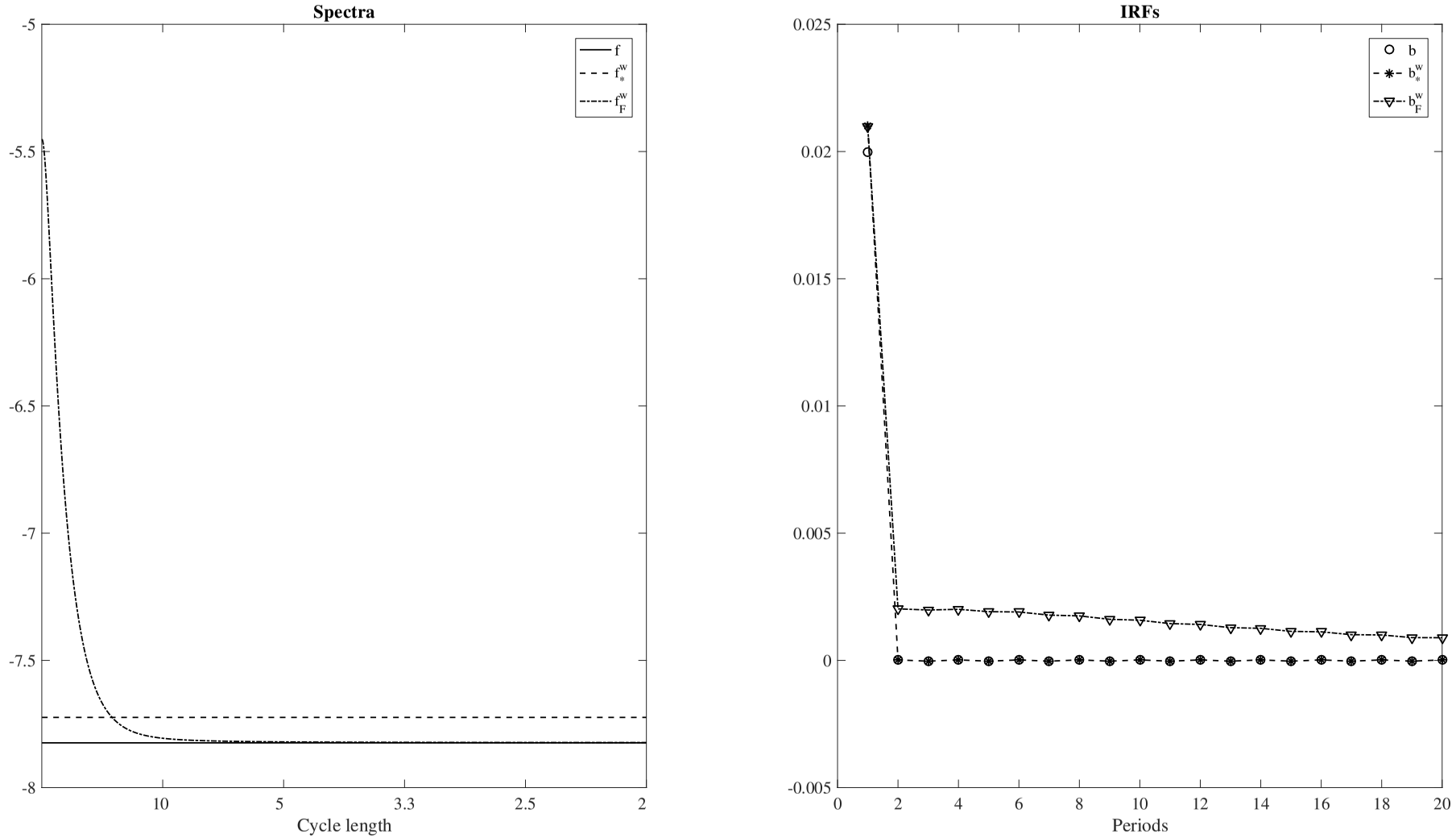
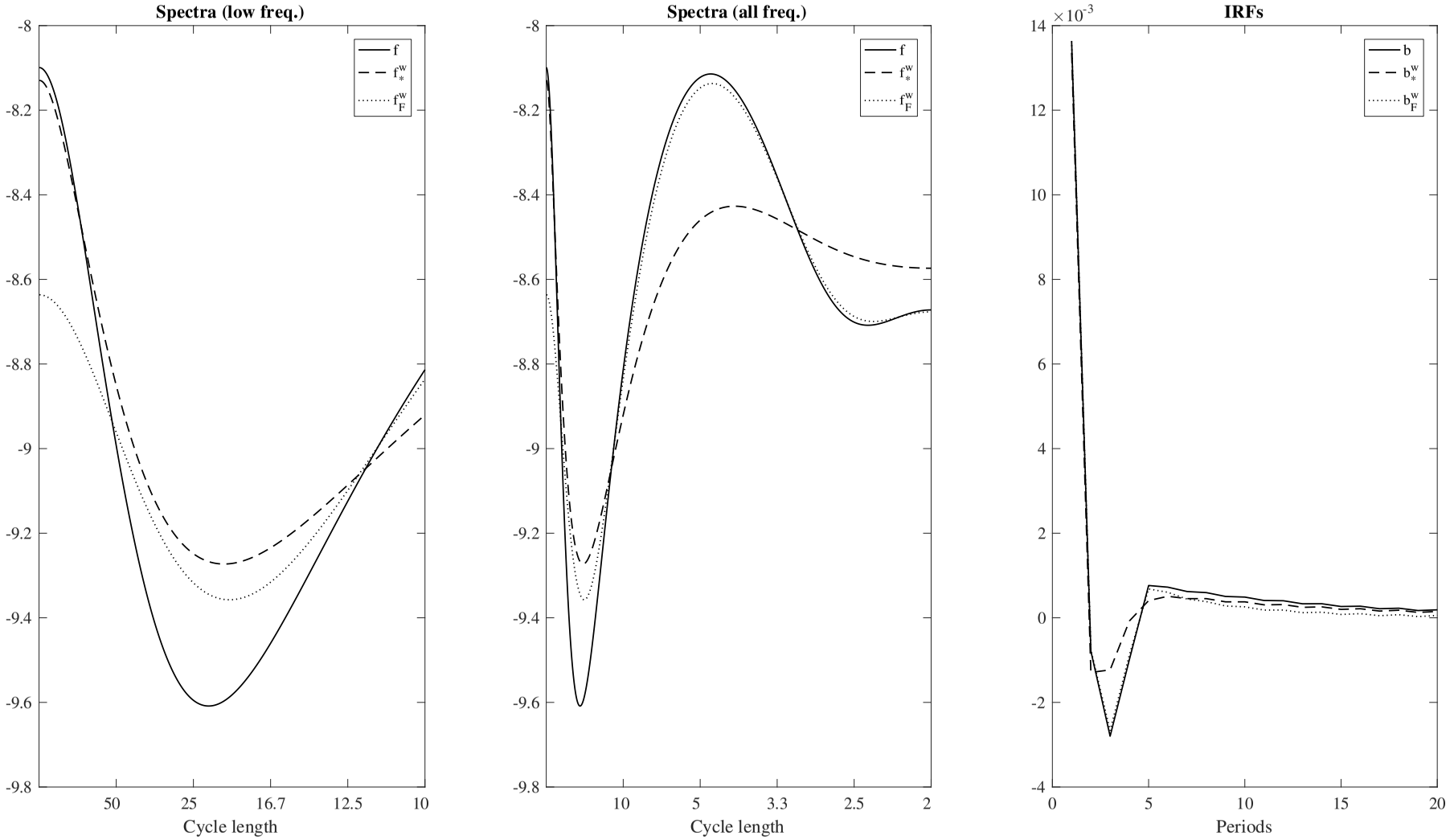


Figure 2: Average estimated log spectra and IRFs for white-noise income



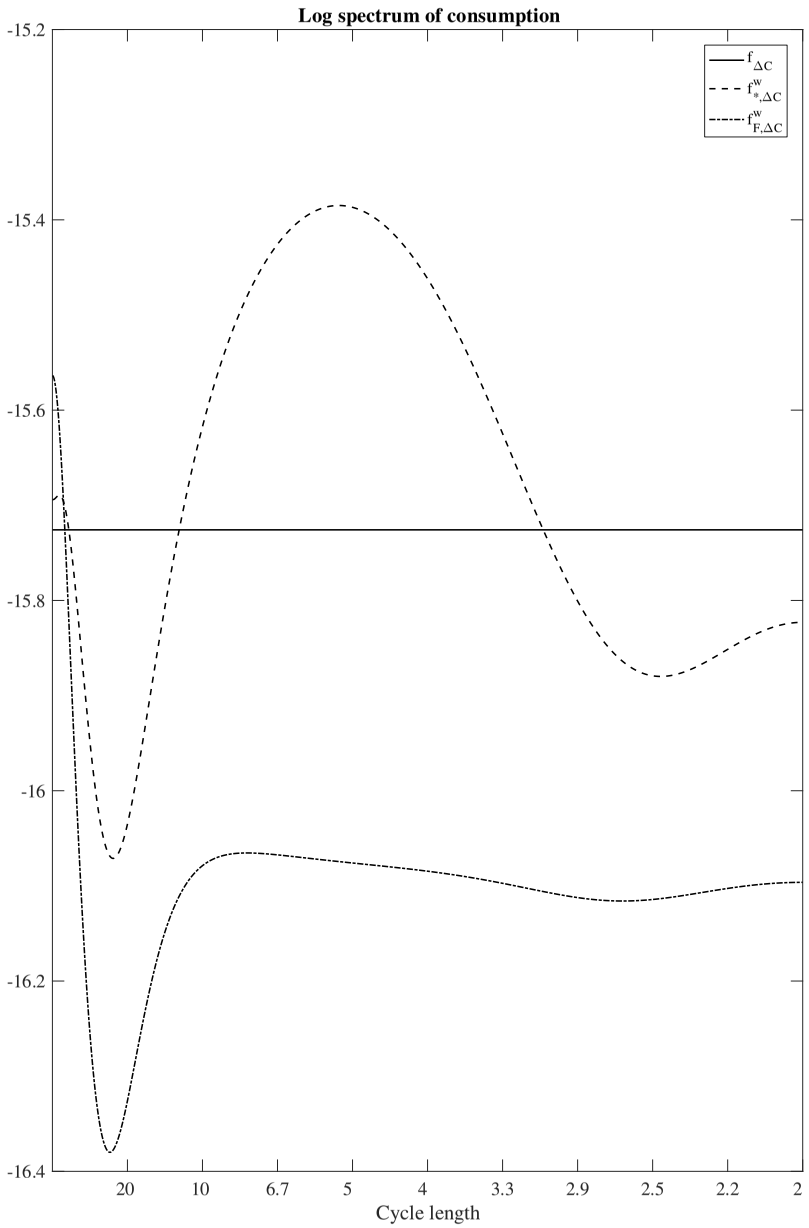
Notes: In the left-hand panel, f is the true log spectrum of income (flat), the line for f_*^w is the average worst-case log spectrum under the optimal information policy, and the line for f_F^w is the average worse-case log spectrum under the statistical benchmark that yields equally precise signals at all frequencies. The right-hand panel plots the impulse response functions (Wold moving average representations) for income corresponding to the three log spectra.

Figure 3: Average estimated log spectra and IRFs with transitory and persistent components in income



Notes: The middle and left-hand panel correspond to the left-hand panel in figure 2, except for a different value for the true spectrum, f. The right-hand panel here corresponds to the right-hand panel in figure 2, but for this alternative example with an income process that has both persistent and transitory components.

Figure 4: Behavior of consumption with permanent and transitory income fluctuations



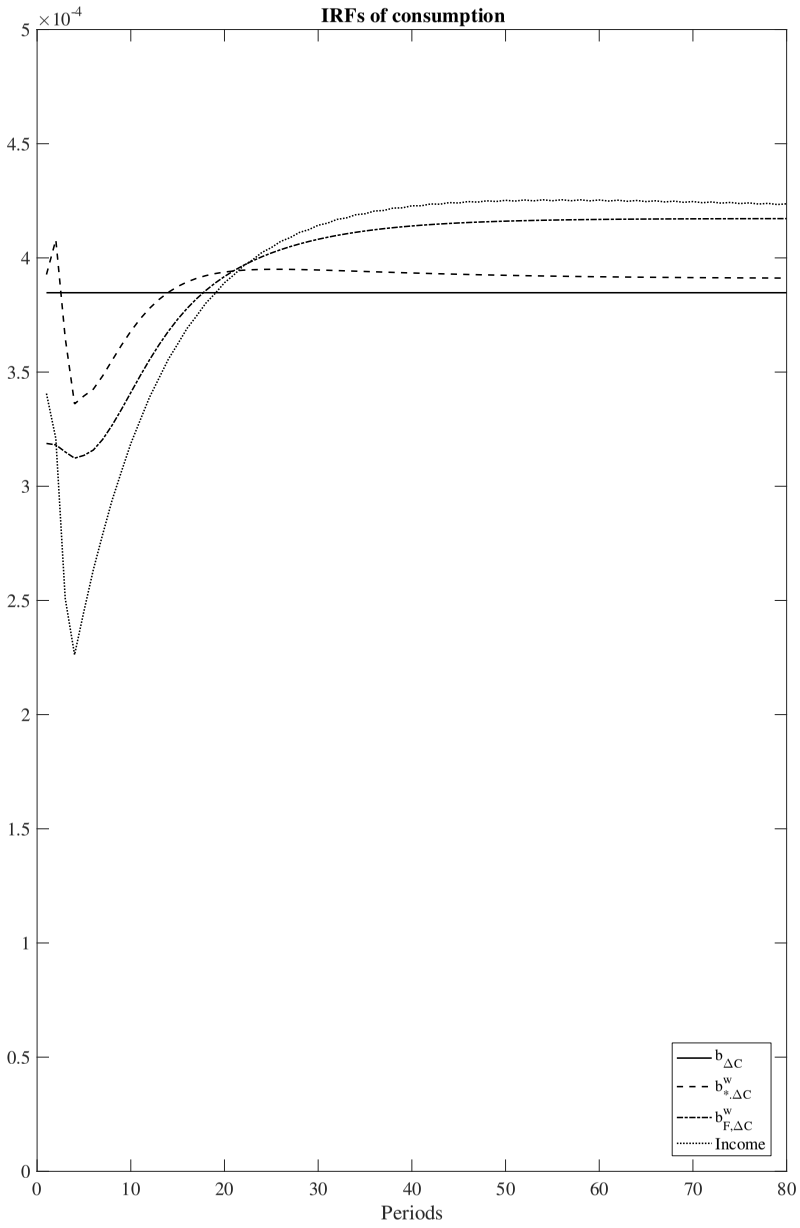
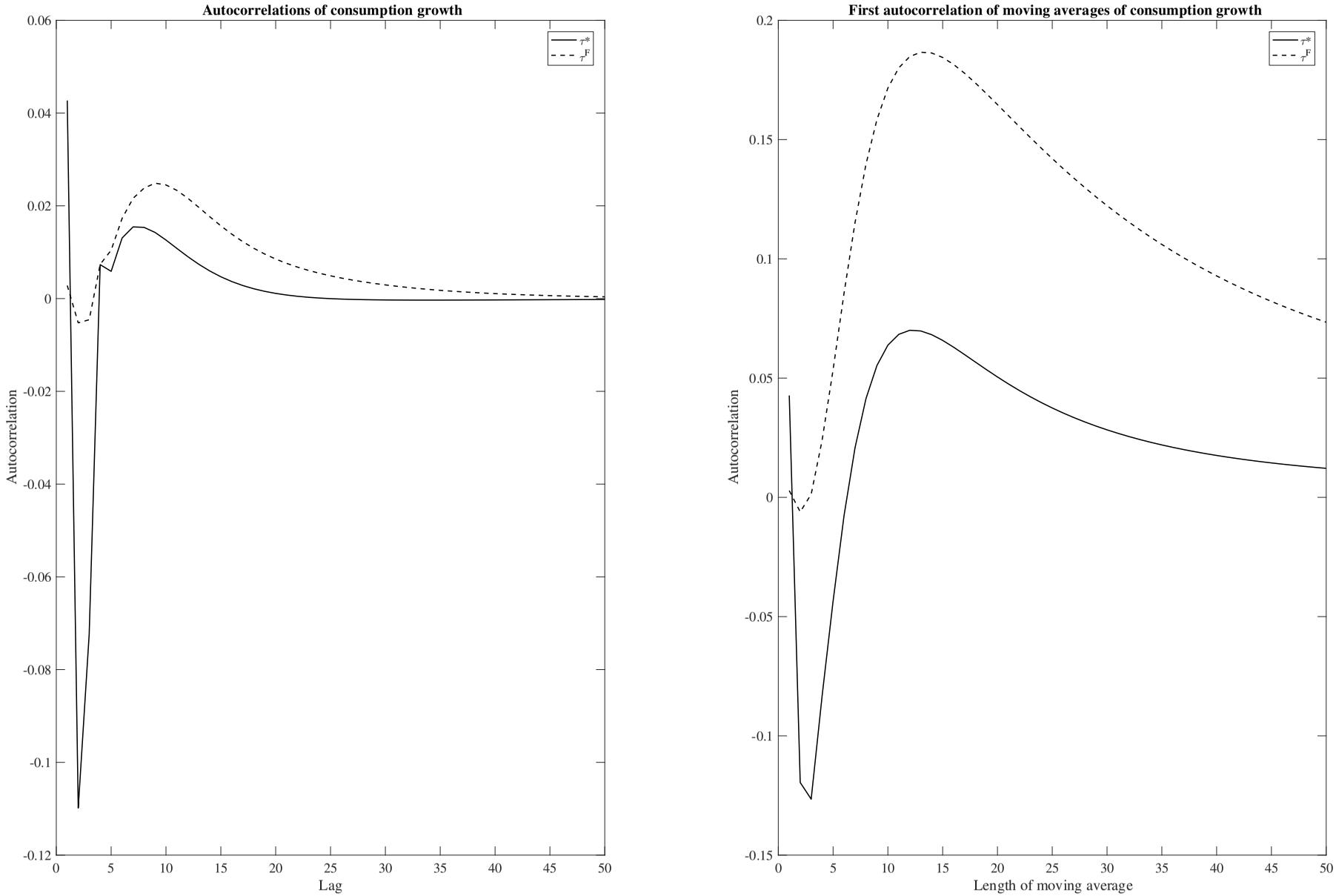
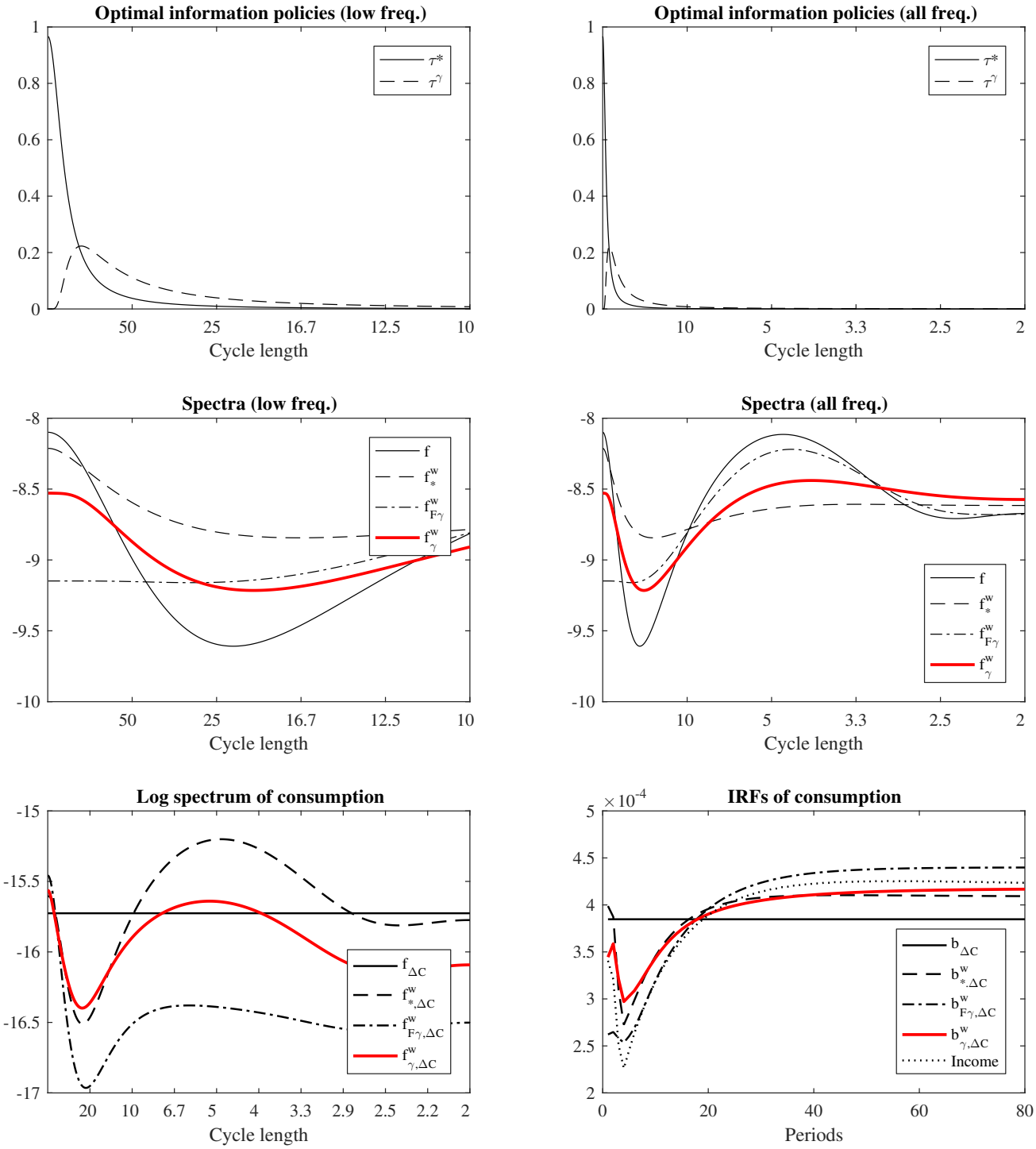


Figure 5: Persistence of consumption growth with transitory and persistent components in income



Notes: The left-hand panel plots the autocorrelations of consumption growth, $corr(\Delta C_t, \Delta C_{t-1})$. The right-hand panel plots the autocorrelation of moving averages, $corr\left(\sum_{j=0}^{n-1} \Delta C_{t+j}, \sum_{j=0}^{n-1} \Delta C_{t-n+j}\right)$, where n varies along the x-axis.

Figure 6: Effects of information cost varying across frequencies



Notes: The middle panels of this figure replicate the left and middle panels in figure 3, but using the optimal information policy when information costs vary across frequencies, denoted τ^{δ}. The bottom panels replicate figure 4. f^w_{δ} denotes the average worst-case log spectrum under that information policy, whereas f^w_* continues to denote the average worst-case spectrum under the optimal policy in the case where information costs are constant across frequencies. $f^w_{F\delta}$ is the average worst-case spectrum under the policy that allocates equal attention cost to each frequency. $f^w_{\delta,\Delta C}$ and $f^w_{F\delta,\Delta C}$ are the spectra for consumption growth associated with those worst-case models.

(note that the probability measure for the expectation operator here is arbitrary) which implies

$$-\alpha^{-1}\log E_t\left[(1-\beta)\sum_{j=0}^{\infty}\beta^j\exp\left(-\alpha C_{t+j}\right)\right] = -\alpha^{-1}\log\frac{(1-\beta)}{1-R}+(R-1)W_{t-1}$$
$$-\frac{\alpha}{2}R^{-1}\left(1-R^{-1}\right)\hat{b}\left(R^{-1}\right)^2+z(L)\hat{\varepsilon}_t-\alpha^{-1}\frac{\log\beta R}{R-1} \quad (42)$$

The result in the text then immediately follows.

B Finding the worst-case spectrum (proposition 1)

Nature chooses $\left\{\hat{f}(\omega_j)\right\}$ to solve

$$\{f^w(\omega_j)\} = \arg\max_{\{\hat{f}(\omega_j)\}} \sum_{j=1}^{n} Z(\omega_j)\hat{f}(\omega_j)d\omega - \frac{\psi^{-1}}{2}\sum_{j=1}^{n}\left(x(\omega_j)-\hat{f}(\omega_j)\right)^2\tau(\omega_j)d\omega \quad (43)$$
$$-\frac{\psi^{-1}}{2}\lambda\sum_{j=2}^{n}\left(\frac{\hat{f}(\omega_j)-\hat{f}(\omega_{j-1})}{d\omega}\right)^2 d\omega. \quad (44)$$

The first-order conditions for interior points ($1<j<n$) are

$$0 = Z(\omega_j)+\psi^{-1}\left(x(\omega_j)-f^w(\omega_j)\right)\tau(\omega)+\frac{\psi^{-1}\lambda}{d\omega}\left(\left(\frac{f^w(\omega_{j+1})-f^w(\omega_j)}{d\omega}\right)-\left(\frac{f^w(\omega_j)-f^w(\omega_{j-1})}{d\omega}\right)\right).$$

At the boundaries they are

$$0 = Z(\omega_1)+\psi^{-1}\left(x(\omega_1)-f^w(\omega_1)\right)\tau(\omega_1)+\psi^{-1}\lambda\frac{f^w(\omega_2)-f^w(\omega_1)}{d\omega^2} \quad (45)$$
$$0 = Z(\omega_n)+\psi^{-1}\left(x(\omega_n)-f^w(\omega_n)\right)\tau(\omega_n)-\psi^{-1}\lambda\frac{f^w(\omega_n)-f^w(\omega_{n-1})}{d\omega^2}. \quad (46)$$

We define here vectors containing the various objects at the frequencies ω_j using variables with no subscript. For example, $\tau \equiv [\tau(\omega_1),\tau(\omega_2),...,\tau(\omega_n)]'$. We can then write the first-order conditions as

$$0 = Z+\psi^{-1}diag(\tau)(x-f^w)+\psi^{-1}\lambda Df^w, \quad (47)$$

where $diag(\tau)$ is a matrix with τ on the diagonal and zero elsewhere and D is a differencing matrix:

$$D \equiv \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & & \\ 0 & 1 & -2 & 1 & & \vdots \\ \vdots & & & \ddots & & 0 \\ & & 0 & 1 & -2 & 1 \\ 0 & \cdots & 0 & 0 & 1 & -1 \end{bmatrix} d\omega^{-2}. \tag{48}$$

The second-order condition is that

$$-diag(\tau) + \lambda D \tag{49}$$

is negative definite, i.e. that all of its eigenvalues are negative.

The solution to nature's optimization problem is then obtained by directly solving (47):

$$\begin{aligned} f^w &= (diag(\tau) - \lambda D)^{-1} (\psi Z + diag(\tau) x) && (50) \\ &= \left(I - \lambda diag\left(\tau^{-1}\right) D\right)^{-1} \left(\psi diag\left(\tau^{-1}\right) Z + x\right), && (51) \end{aligned}$$

where τ^{-1} here is an elementwise inverse of the vector τ. Since this is a linear problem, the solution is unique as long as the matrix inverse exists.

C Proposition 2

Consider a total derivative of (47) with respect to τ' at the point $x = \bar{f}$:

$$0 = \psi^{-1} diag\left(\bar{f} - f^w\right) - \psi^{-1} diag(\tau) \frac{dE[f^w]}{d\tau'} + \psi^{-1} \lambda D \frac{dE[f^w]}{d\tau'}. \tag{52}$$

We can then solve for $\frac{dE[f^w]}{d\tau'}$:

$$\frac{dE[f^w]}{d\tau'} = \left(\lambda D - diag\left(\tau'\right)\right)^{-1} diag\left(E[f^w] - \bar{f}\right) \tag{53}$$

Now the objective is to minimize

$$\begin{aligned} \{\tau^*(\omega_j)\} &= \arg\min_{\{\tau(\omega_j)\}} \log E\left[b^w \left(R^{-1}\right)^2\right] + \theta \sum_j \tau(\omega_j) d\omega && (54) \\ &= \arg\min_{\{\tau(\omega_j)\}} Z' E[f^w] d\omega + \theta \sum_j \tau(\omega_j) d\omega. && (55) \end{aligned}$$

The first-order condition for that problem is

$$0 = Z' \frac{dE[f^w]}{d\tau'} + \theta 1_{1\times n}, \tag{56}$$

where $1_{1\times n}$ is a $1 \times n$ vector of ones. Inserting the formula for $\frac{df^w}{d\tau'}$ yields

$$0 = Z' \left(\lambda D - diag\left(\tau^{*\prime}\right)\right)^{-1} diag\left(E\left[f^w\right] - \bar{f}\right) + \theta 1_{1\times n} \tag{57}$$

$$Z' = -\theta 1_{1\times n} diag\left(E\left[f^w\right] - \bar{f}\right)^{-1} \left(\lambda D - diag\left(\tau^{*\prime}\right)\right). \tag{58}$$

Now we conjecture that $E\left[f^w\right] - \bar{f}$ is equal to a constant c multiplied by a column of ones. We then have

$$\begin{aligned} Z' &= -\theta c^{-1} 1_{1\times n} \left(\lambda D - diag\left(\tau^{*\prime}\right)\right) && (59)\\ &= \theta c^{-1} \tau^{*\prime}, && (60) \end{aligned}$$

where the second line uses the fact that $1_{1\times n} D = 0_{1\times n}$ since the columns of D sum to zero.

In order to confirm that result, we must now show that when $Z = \theta c^{-1}\tau^*$, $E\left[f^w\right] - \bar{f} = c1_{n\times 1}$. Inserting $Z = \theta c^{-1}\tau^*$ into (47) yields

$$0 = \theta c^{-1}\tau^* + \psi^{-1} diag\left(\tau^*\right)\left(\bar{f} - E\left[f^w\right]\right) + \psi^{-1}\lambda D E\left[f^w\right]. \tag{61}$$

In order for it to be the case that $E\left[f^w\right] - \bar{f} = c1_{n\times 1}$, we must have

$$\begin{aligned} 0 &= \theta c^{-1}\tau^* - \psi^{-1} diag\left(\tau^*\right) 1_{n\times 1} c + \psi^{-1}\lambda D 1_{n\times 1}\left(\bar{f} + c\right) && (62)\\ &= \theta c^{-1}\tau^* - \psi^{-1}\tau^* c. && (63) \end{aligned}$$

where the second line uses the fact that $D1_{n\times 1} = 0_{n\times 1}$. This is solved by

$$\sqrt{\theta\psi} = c \tag{64}$$

$$\begin{aligned} Z &= \theta c^{-1}\tau^* && (65)\\ \tau^* &= (\theta/\psi)^{-1/2} Z. && (66) \end{aligned}$$

We can then plug the value of τ^* into the equation for $E\left[f^w\right]$:

$$\begin{aligned} E\left[f^w\right] &= \left(I - \lambda diag\left(\tau^{*-1}\right) D\right)^{-1} \left(\psi diag\left(\left(\theta/\psi\right)^{1/2} Z^{-1}\right) Z + x\right) && (67)\\ &= \left(I - \lambda diag\left(\tau^{*-1}\right) D\right)^{-1} \left(1_{n\times 1}\theta^{1/2}\psi^{1/2} + x\right), && (68) \end{aligned}$$

where, as with τ^{-1}, Z^{-1} is an elementwise inverse of the vector Z. It follows that

$$\begin{aligned} E\left[f^w\right] &= \left(I - \lambda diag\left(\tau^{*-1}\right) D\right)^{-1} \left(1_{n\times 1}\theta^{1/2}\psi^{1/2} + \bar{f}\right) && (69)\\ &= 1_{n\times 1}\theta^{1/2}\psi^{1/2} + \bar{f}, && (70) \end{aligned}$$

where the last line follows from the fact that the rows of $\left(I-\lambda diag\left(\tau^{*-1}\right)D\right)^{-1}$ sum to 1. To see why, note that

$$\left(I-\lambda diag\left(\tau^{*-1}\right)D\right)^{-1}=I+\lambda diag\left(\tau^{*-1}\right)D+\left(\lambda diag\left(\tau^{*-1}\right)D\right)^{2}+... \tag{71}$$

The rows of $\lambda diag\left(\tau^{*-1}\right)D$ sum to zero, meaning that $1_{n\times 1}$ is an eigenvector with eigenvalue zero. When a matrix is raised to a power, its eigenvectors are unchanged and its eigenvalues are raised to the same power, meaning that $1_{n\times 1}$ remains an eigenvector with 0 the associated eigenvalue, and the rows sum to zero. Since the rows of I sum to 1, the rows of $\left(I-\lambda diag\left(\tau^{*-1}\right)D\right)^{-1}$ then do also.

C.1 Bias of $f^{w}(\omega;x,\tau)$

From above, the solution for the vector f^{w} is

$$f^{w}(x,\tau) = \left(I-\lambda diag\left(\tau^{-1}\right)D\right)^{-1}\left(\psi diag\left(\tau^{-1}\right)Z+x\right) \tag{72}$$

$$f^{w}(x,\tau) = \left(I+\sum_{j=1}^{\infty}\left(\lambda diag\left(\tau^{-1}\right)D\right)^{j}\right)\left(\psi diag\left(\tau^{-1}\right)Z+x\right) \tag{73}$$

$$f^{w}(x,\tau)-\psi diag\left(\tau^{-1}\right)Z-x = \left(\sum_{j=1}^{\infty}\left(\lambda diag\left(\tau^{-1}\right)D\right)^{j}\right)\left(\psi diag\left(\tau^{-1}\right)Z+x\right). \tag{74}$$

Now scale τ^{-1} by c and divide both sides by c

$$c^{-1}f^{w}(x,\tau/c)-\psi diag\left(\tau^{-1}\right)Z-c^{-1}x=\begin{pmatrix}\lambda diag\left(\tau^{-1}\right)D\\ +\sum_{j=2}^{\infty}c^{j-1}\left(\lambda diag\left(\tau^{-1}\right)D\right)^{j}\end{pmatrix}\left(c\psi diag\left(\tau^{-1}\right)Z+x\right). \tag{75}$$

Since both sides are linear in x, we can take take the expectation and then the limit as $c\to 0$ to yield

$$\lim_{c\to 0}\frac{E\left[f^{w}(x,\tau/c)\right]-f}{c}=\psi diag\left(\tau^{-1}\right)Z+\lambda diag\left(\tau^{-1}\right)Df. \tag{76}$$

In the limit as $n\to\infty$, Df becomes f''.

C.2 Alternative information cost specifications

When $\lambda=0$, the formula for f^{w} as a function of τ becomes

$$f^{w}(\omega_j)=\psi\tau(\omega_j)^{-1}Z(\omega_j)+x(\omega_j) \tag{77}$$

and the first-order condition for the optimization of $\tau(\omega_j)$ is in the case with arbitrary $\gamma(\omega_j)$

$$0=-\psi\tau(\omega_j)^{-2}Z(\omega_j)^{2}+\theta\gamma(\omega_j) \tag{78}$$

The result from the text then follows immediately. For the entropy information cost, the first-order condition is

$$0 = -\psi\tau\left(\omega_j\right)^{-2} Z\left(\omega_j\right)^2 + \theta\tau\left(\omega_j\right)^{-1} \tag{79}$$

which again immediately yields the desired result.

Online appendix for "Directed Attention and Nonparametric Learning"

Ian Dew-Becker and Charles G. Nathanson

March 10, 2019

1 The behavior of consumption

The Euler equation conditional on a model $\hat{b}$ is

$$1 = E_t\left[\beta \exp\left(-a\Delta C_{t+1}\right) R \mid \hat{b}\right] \tag{1}$$

It is then straightforward to confirm that is solved by

$$\begin{aligned} C_t &= (R-1) W_{t-1} + Z_t - (R-1)^{-1} \alpha^{-1} \log(\beta R) && (2)\\ W_t &= W_{t-1} + Y_t - Z_t + (R-1)^{-1} \alpha^{-1} \log(\beta R), && (3) \end{aligned}$$

where

$$Z_t = \left(1 - R^{-1}\right) Y_t - \frac{1}{\alpha} R^{-1} \log E_t \exp\left(-\alpha Z_{t+1}\right). \tag{4}$$

We then have

$$\Delta C_{t+1} = (R-1) Y_t + Z_{t+1} - R Z_t + \alpha^{-1} \log(\beta R). \tag{5}$$

Now define H as follows:

$$\begin{aligned} H_t &\equiv Z_t - \left(1 - R^{-1}\right) Y_t && (6)\\ H_t &= -\frac{1}{\alpha} R^{-1} \log E_t \exp\left(-\alpha\left(H_{t+1} + \left(1 - R^{-1}\right) Y_{t+1}\right)\right). && (7) \end{aligned}$$

This definition yields

$$\begin{aligned} \Delta C_{t+1} &= R\left(\left(1 - R^{-1}\right) Y_t - Z_t\right) + Z_{t+1} + \alpha^{-1} \log(\beta R) && (8)\\ &= H_{t+1} - R H_t + \left(1 - R^{-1}\right) Y_{t+1} + \alpha^{-1} \log(\beta R). && (9) \end{aligned}$$

Guessing $H_t = \bar{h} + h(L)\varepsilon_t$, we have the recursion

$$\begin{aligned} \bar{h} + h(L)\varepsilon_t &= -\frac{1}{\alpha} R^{-1} \log E_t\left[\exp\left(-\alpha\left(\bar{h} + h(L)\hat{\varepsilon}_{t+1} + \left(1 - R^{-1}\right)\hat{b}(L)\hat{\varepsilon}_{t+1}\right)\right) \mid f^w\right] && (10)\\ &= R^{-1}\left(\bar{h} + \sum_{j=1}^{\infty}\left(h_j + \left(1 - R^{-1}\right)\hat{b}_j\right)\hat{\varepsilon}_{t+1-j}\right) - R^{-1}\frac{\alpha}{2}\left(h_0 + \left(1 - R^{-1}\right)\hat{b}_0\right)^2 && (11) \end{aligned}$$

where $\hat{\varepsilon}_t \equiv \hat{b}(L)^{-1} Y_t$. The solution is

$$\begin{aligned} h_j &= R^{-1} h_{j+1} + R^{-1}\left(1 - R^{-1}\right)\hat{b}_{j+1} && (12)\\ \bar{h} &= -\frac{R^{-1}}{1 - R^{-1}}\frac{\alpha}{2}\left(h_0 + \left(1 - R^{-1}\right)\hat{b}_0\right)^2 && (13)\\ &= -R^{-1}\frac{\alpha}{2}\left(1 - R^{-1}\right)\hat{b}\left(R^{-1}\right)^2 && (14) \end{aligned}$$

Now we can insert the formulas for the various objects:

$$\begin{aligned}
\Delta C_t &= \left(1-R^{-1}\right) b\left(L\right) \varepsilon_{t+1} + \left(1-R\right) \bar{h} + h_0 \hat{\varepsilon}_{t+1} + \sum_{j=0}^{\infty} \left(h_{j+1} - R h_j\right) \hat{\varepsilon}_{t-j} + \alpha^{-1} \log \beta R && (15)\\
&= \left(1-R^{-1}\right) b\left(L\right) \varepsilon_{t+1} + \left(1-R\right) \bar{h} + \left(1-R^{-1}\right) \left(\hat{b}\left(R^{-1}\right) - \hat{b}_0\right) \hat{\varepsilon}_{t+1} && (16)\\
&\quad - \sum_{j=0}^{\infty} \left(1-R^{-1}\right) \hat{b}_{j+1} \hat{\varepsilon}_{t-j} + \alpha^{-1} \log \beta R && (17)\\
&= \left(1-R^{-1}\right) b\left(L\right) \varepsilon_{t+1} + \left(1-R\right) \bar{h} + \left(1-R^{-1}\right) \left(\hat{b}\left(R^{-1}\right) - \hat{b}_0\right) \frac{b\left(L\right)}{\hat{b}\left(L\right)} \varepsilon_{t+1} && (18)\\
&\quad - \sum_{j=0}^{\infty} \left(1-R^{-1}\right) \hat{b}_{j+1} \frac{b\left(L\right)}{\hat{b}\left(L\right)} \varepsilon_{t-j} + \alpha^{-1} \log \beta R && (19)\\
&= \left(1-R^{-1}\right) b\left(L\right) \varepsilon_{t+1} + \left(1-R\right) \bar{h} + \left(1-R^{-1}\right) \hat{b}\left(R^{-1}\right) \frac{b\left(L\right)}{\hat{b}\left(L\right)} \varepsilon_{t+1} && (20)\\
&\quad - \left(1-R^{-1}\right) \hat{b}\left(L\right) \frac{b\left(L\right)}{\hat{b}\left(L\right)} \varepsilon_{t+1} + \alpha^{-1} \log \beta R && (21)\\
&= \left(1-R^{-1}\right) \hat{b}\left(R^{-1}\right) \frac{b\left(L\right)}{\hat{b}\left(L\right)} \varepsilon_{t+1} + \left(1-R\right) \bar{h} + \alpha^{-1} \log \beta R. && (22)
\end{aligned}$$

So consumption growth is equal to a constant plus $\left(1-R^{-1}\right) \hat{b}\left(R^{-1}\right) \frac{b(L)}{\hat{b}(L)} \varepsilon_{t+1}$. The dynamic behavior of consumption growth is therefore determined by $\hat{b}\left(R^{-1}\right) \frac{b(L)}{\hat{b}(L)}$. To help with the intuition, note that

$$\frac{b\left(L\right)}{\hat{b}\left(L\right)} \varepsilon_{t+1} = \hat{\varepsilon}_{t+1} \tag{23}$$

so this simply says that people follow the usual consumption rule but applied to their filtered innovations, which may not be the true innovations.

The spectral density of consumption growth is

$$\hat{f}_{\Delta C}\left(\omega\right) = \hat{b}\left(R^{-1}\right)^2 \frac{f\left(\omega\right)}{\hat{f}\left(\omega\right)}. \tag{24}$$

2 Epstein–Zin preferences

Suppose people have preferences of the form

$$v_t = \left(1-\beta\right) c_t + \frac{\beta}{1-\alpha} \log E_t \left[\exp\left(\left(1-\alpha\right) v_{t+1}\right)\right] \tag{25}$$

where $c_t = \log C_t$. They face the budget constraint

$$W_{t+1} = R_{t+1}\left(W_t - C_t\right) \tag{26}$$

Returns follow the process

$$r_{t+1} = \log R_{t+1} = \bar{r} + b\left(L\right) \varepsilon_{t+1} \tag{27}$$

lower-case letters from here on denote logs.

Since people have a unit elasticity of intertemporal substitution, the consumption-wealth ratio will be constant. We write

$$c_t = \bar{c} + w_t \tag{28}$$

The budget constraint can be rewritten as

$$\Delta w_{t+1} = r_{t+1} + \log\left(1 - \exp\left(\bar{c}\right)\right) \tag{29}$$

where Δ is the first-difference operator.

For consumption growth, we then have

$$\begin{aligned} \Delta c_{t+1} &= \Delta w_{t+1} \\ &= r_{t+1} + \log\left(1 - \exp\left(\bar{c}\right)\right) \end{aligned}$$

and we guess that lifetime utility is

$$v_t = \bar{v} + c_t + v\left(L\right)\varepsilon_t \tag{30}$$

We can confirm this guess,

$$\bar{v} + c_t + v\left(L\right)\varepsilon_t = \left(1-\beta\right)c_t + \frac{\beta}{1-\alpha}\log E_t\left[\exp\left(\left(1-\alpha\right)\left(\bar{v} + c_{t+1} + v\left(L\right)\varepsilon_{t+1}\right)\right)\right] \tag{31}$$

$$\bar{v} + v\left(L\right)\varepsilon_t = \frac{\beta}{1-\alpha}\log E_t\left[\exp\left(\left(1-\alpha\right)\left(\bar{v} + \log\left(1-\exp\left(\bar{c}\right)\right) + \bar{r} + b\left(L\right)\varepsilon_{t+1} + v\left(L\right)\varepsilon_{t+1}\right)\right)\right] \tag{32}$$

$$= \beta\left(\bar{v} + \log\left(1-\exp\left(\bar{c}\right)\right) + \bar{r} + b_+\left(L\right)\varepsilon_{t+1} + v_+\left(L\right)\varepsilon_{t+1}\right) + \beta\frac{1-\alpha}{2}\left(b_0 + v_0\right)^2\sigma^2 \tag{33}$$

where v_+ and b_+ denote the lag polynomials with the constants (b_0 and v_0) removed and the coefficients in the polynomials are denoted b_j and v_j.

Matching coefficients yields

$$\bar{v} = \frac{\beta}{1-\beta}\left(\log\left(1-\exp\left(\bar{c}\right)\right) + \bar{r} + \frac{1-\alpha}{2}\left(b_0 + v_0\right)^2\sigma^2\right) \tag{34}$$

$$v_j = \beta\left(v_{j+1} + b_{j+1}\right) \tag{35}$$

$$\Rightarrow \quad v_0 + b_0 = b\left(\beta\right) \tag{36}$$

So we have

$$\bar{v} = \frac{\beta}{1-\beta}\left(E\left[\Delta c\right] + \frac{1-\alpha}{2}b\left(\beta\right)^2\sigma_\varepsilon^2\right) \tag{37}$$

Next we insert this into the Euler equation along with the expression for consumption growth

$$\begin{aligned} 1 &= E_t\left[\beta\frac{\exp\left(\left(1-\alpha\right)v_{t+1} - \Delta c_{t+1} + r_{t+1}\right)}{E_t\left[\exp\left(\left(1-\alpha\right)v_{t+1}\right)\right]}\right] \\ \bar{c} &= \log\left(1-\beta\right) \end{aligned}$$

finally yielding

$$\bar{v} = \frac{\beta}{1-\beta}\left(\log\left(\beta\right) + \bar{r} + \frac{1-\alpha}{2}b\left(\beta\right)^2\sigma^2\right) \tag{38}$$

The key result here is that lifetime utility depends on $b\left(\beta\right)^2\sigma_\varepsilon^2$, which is the same term as in the main text. Note also that log utility is the special case of the above in which $\alpha = 1$. In that case, agents are indifferent to return risk.

3 Prior on smoothness in terms of cycle length

The main analysis studies the spectrum in the frequency domain, with a prior on smoothness that is equally strong at all frequencies. This section considers an alternative specification where the analysis is in terms of cycle length and shows that the optimal information acquisition policy remains unchanged, even though lower-frequency fluctuations become more difficult to learn about.

Recall from the main text that for a fluctuation at a frequency ω, the length of the associated cycle is $\zeta = 2\pi/\omega$. It is obviously possible through a simple change of variables to write the entire model in terms of cycles instead of frequencies.

The key difference in this section from the main text is that we assume that agents have a prior on the smoothness of the spectrum in terms of cycles, rather than frequencies. Note that in the frequency domain, the upper half of the range $[0, \pi]$ is associated with cycles lasting four or fewer periods, whereas the lower half of the range is associated with all longer cycles. In economic terms, we might think there is potentially much more interesting variation in the model in the range of cycles lasting longer than four periods. Put another way, any cycle lasting less than, say, four quarters, could be said to be "high", with little meaningful to distinguish them, whereas cycles lasting longer than four quarters could include business cycles, medium-frequency trends, and long-term growth rates. It might be more natural, then, for the agent to have a prior on the smoothness of the spectrum written in terms of cycles than frequencies.

Formally, define

$$\tilde{f}(\zeta) \equiv f\left(\frac{2\pi}{\zeta}\right) \tag{39}$$

We now say that the agent has a prior that restricts the total squared variation in $\tilde{f}$, which would be

$$\begin{aligned} \int_2^\infty \tilde{f}'(\zeta)^2 d\zeta &= \int_0^\pi \tilde{f}'\left(\frac{2\pi}{\omega}\right)^2 \frac{2\pi}{\omega^2} d\omega & (40) \\ &= \int_0^\pi f'\left(\frac{2\pi}{\omega}\right)^2 \omega^2 d\omega & (41) \end{aligned}$$

In other words, since the transformation $2\pi/\omega$ stretches the space around the very lowest frequencies, the agent is essentially open to the possibility that the spectrum might be infinitely variable at the very lowest frequencies.

Going back to the discretization used in the main analysis, we write the penalized likelihood in this case as

$$\begin{aligned} P\left(\hat{f} \mid x, \tau\right) &= -\frac{1}{2} d\omega \sum_{j=1}^{n} \left(x(\omega_j) - f(\omega_j)\right)^2 \tau(\omega_j) - \frac{\lambda}{2} \sum_{j=2}^{n} \left(\frac{\hat{f}(\omega_j) - \hat{f}(\omega_{j-1})}{d\omega}\right)^2 \omega_j^2 d\omega & (42) \\ &\quad + \text{constants} & (43) \end{aligned}$$

with the only difference now being the added ω_j^2 in the second summation, showing that the smoothness prior is tighter at high than low frequencies. This can be written in terms of vectors and matrices as

$$\begin{aligned} P\left(\hat{f} \mid x, \tau\right) &= -\frac{1}{2} d\omega \left(x - \hat{f}\right)' diag(\tau) \left(x - \hat{f}\right) - \frac{\lambda}{2} \hat{f}' D_\omega \hat{f} d\omega & (44) \\ &\quad + \text{constants} & (45) \end{aligned}$$

where

$$D_\omega \equiv \begin{bmatrix} -\omega_2^2 & \omega_2^2 & 0 & 0 & \cdots & 0 \\ \omega_2^2 & -\omega_3^2 - \omega_2^2 & \omega_3^2 & 0 & & \\ 0 & \omega_3^2 & -\omega_4^2 - \omega_3^2 & \omega_4^2 & & \vdots \\ \vdots & & & \ddots & & 0 \\ & & 0 & \omega_{n-1}^2 & -\omega_n^2 - \omega_{n-1}^2 & \omega_n^2 \\ 0 & \cdots & 0 & 0 & \omega_n^2 & -\omega_n^2 \end{bmatrix} d\omega^{-2}. \tag{46}$$

3.1 Estimation precision

There is a formal sense in which the change in the smoothness prior makes it more difficult for an agent to learn about low frequencies. Consider the simple estimation problem of choosing $\hat{f}$ to maximize the posterior

probability $P\left(\hat{f} \mid x, \tau\right)$. The point estimate is then

$$f^* \equiv \arg\max_{\hat{f}} P\left(\hat{f} \mid x, \tau\right) \tag{47}$$

$$= \left(diag\left(\tau\right) - \lambda D_\omega\right)^{-1} diag\left(\tau\right) x \tag{48}$$

If we set $\tau = \bar{\tau} 1_{n\times 1}$ for a scalar $\bar{\tau}$, so that the agent has signals with equal precision at all frequencies, we obtain

$$f^* = \left(I - \lambda \bar{\tau}^{-1} D_\omega\right)^{-1} x \tag{49}$$

The variance matrix of f^* is

$$var\left(f^*\right) = \bar{\tau}^{-1}\left(I - \lambda \bar{\tau}^{-1} D_\omega\right)^{-1}\left(I - \lambda \bar{\tau}^{-1} D_\omega\right)^{-1} \tag{50}$$

The variance is straightforward to analyze numerically. Figure A.1 plots the main diagonal of the variance matrix for various values of $\lambda\bar{\tau}^{-1}$. Each case is rescaled so that they are equal for the lowest frequency, illustrating how the variances differ across frequencies. In all cases, the variance of the estimator $f^*\left(\omega\right)$ is lower at high frequencies. So when agents have a weaker prior on smoothness at low than high frequencies (due to the assumption of equal smoothness in terms of cycles), it is more difficult to learn about the spectrum at low frequencies.

3.2 Optimal information policy

It is straightforward to confirm that the optimal information policy, $\tau \propto Z$, is unchanged in this case. The result follows from the fact that the two characteristics of D that are necessary for the main result – that its rows and columns sum to zero – also hold for D_ω.

4 Case with a ceiling on the length of available income histories

The discussion in section 2 notes that our analysis assumes that the agent is able to find income histories in their database that are potentially arbitrarily long. That assumption is model consistent, but it is still valuable to check the model's robustness to it. The main text checks robustness by assuming that the cost of acquiring signals about the spectrum becomes infinite as the frequency approaches zero. This section considers an alternative case where there is a fixed frequency below which agents are completely unable to acquire signals. This corresponds to a case where the database of income histories contains no histories longer than some specific cutoff.

To implement this analysis, we simply assume that the cost of acquiring signals, $\gamma\left(\omega\right)$, becomes infinite for $\omega < \bar{\omega}$. We retain the same calibration as in the main text for the case of an income spectrum with two peaks. We set $\bar{\omega}$ to correspond to cycles lasting 74 years or longer. We choose that length because it corresponds to the length of the post-war US period, which is a common sample used in estimation. It is also similar to typical lifespans (though somewhat longer than working lives, in general).

Figure A.2 replicates figure 6 from the main text, but using the hard cutoff for γ. The top two panels show that the precision of the signals agents acquire in this case has a large spike at low frequencies. That point is just to the right of the frequency cutoff. Intuitively, since agents cannot learn about the very lowest frequencies, they shift the attention that would have gone to those frequencies to the lowest they are able to observe. Their smoothness prior means that the information at $\bar{\omega}$ is still useful for providing information about the spectrum at frequencies below $\bar{\omega}$.

The results in the second and third rows of figure A.2 are very similar to those in figure 6. Specifically, the optimal information policy in this case again yields beliefs about the spectrum for income, along with a spectrum for consumption growth, that lie between those obtained under the optimum with constant information costs (τ^*) and those for the equal-cost policy, $\tau^{F\gamma}$. Here, $\tau^{F\gamma}$ is equal to zero for $\omega < \bar{\omega}$ and a constant otherwise, since all $\omega \geq \bar{\omega}$ have equal cost.

Figure A.2 therefore shows that optimal information choice continues to cause agents to learn more about the lowest frequencies than they would under a purely statistical policy. This causes them to make

comparatively larger mistakes at middle and high frequencies, but smaller mistakes at low frequencies. As in figure 6, that result is weaker than in our baseline case with constant costs across all frequencies, but the same intuition continues to go through.

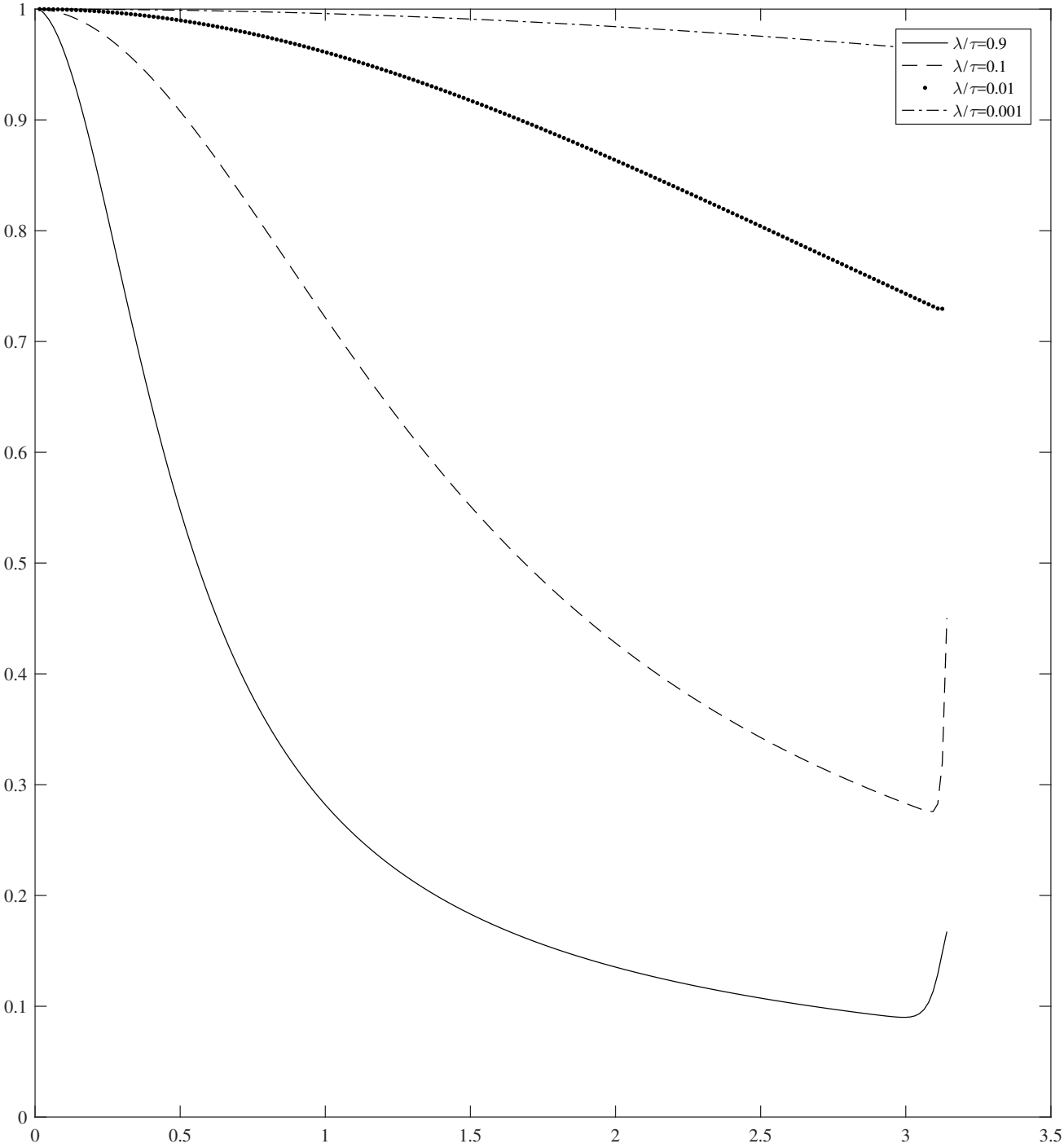
5 Small-sample simulations for the periodogram distribution

Our analysis uses the standard time series result that the periodogram is exponentially distributed around the true spectrum with errors that are uncorrelated across frequencies. This section checks the accuracy of that result in small-sample simulations of the two-peak spectrum used for most of the analysis. Since that spectrum has a peak at low frequencies, it accounts for the potential concern that the results might not be accurate in the presence of strong persistence.

We simulate 100-year histories of the income process. For each history, we calculate the periodogram at the fundamental Fourier frequencies. Figure A.3 plots the standard deviation of the log periodogram across 1,000,000 simulations at the 49 Fourier frequencies ($2\pi j/100$ for $j \in \{1, 2, ..., 49\}$; the other frequencies mirror the first 49, and the analysis ignores frequencies 0 and π, which have slightly different distributions). The horizontal line is set at $\pi/\sqrt{6}$, which is the theoretical standard deviation. The simulated standard deviations are tightly clustered around the theoretical standard deviation.

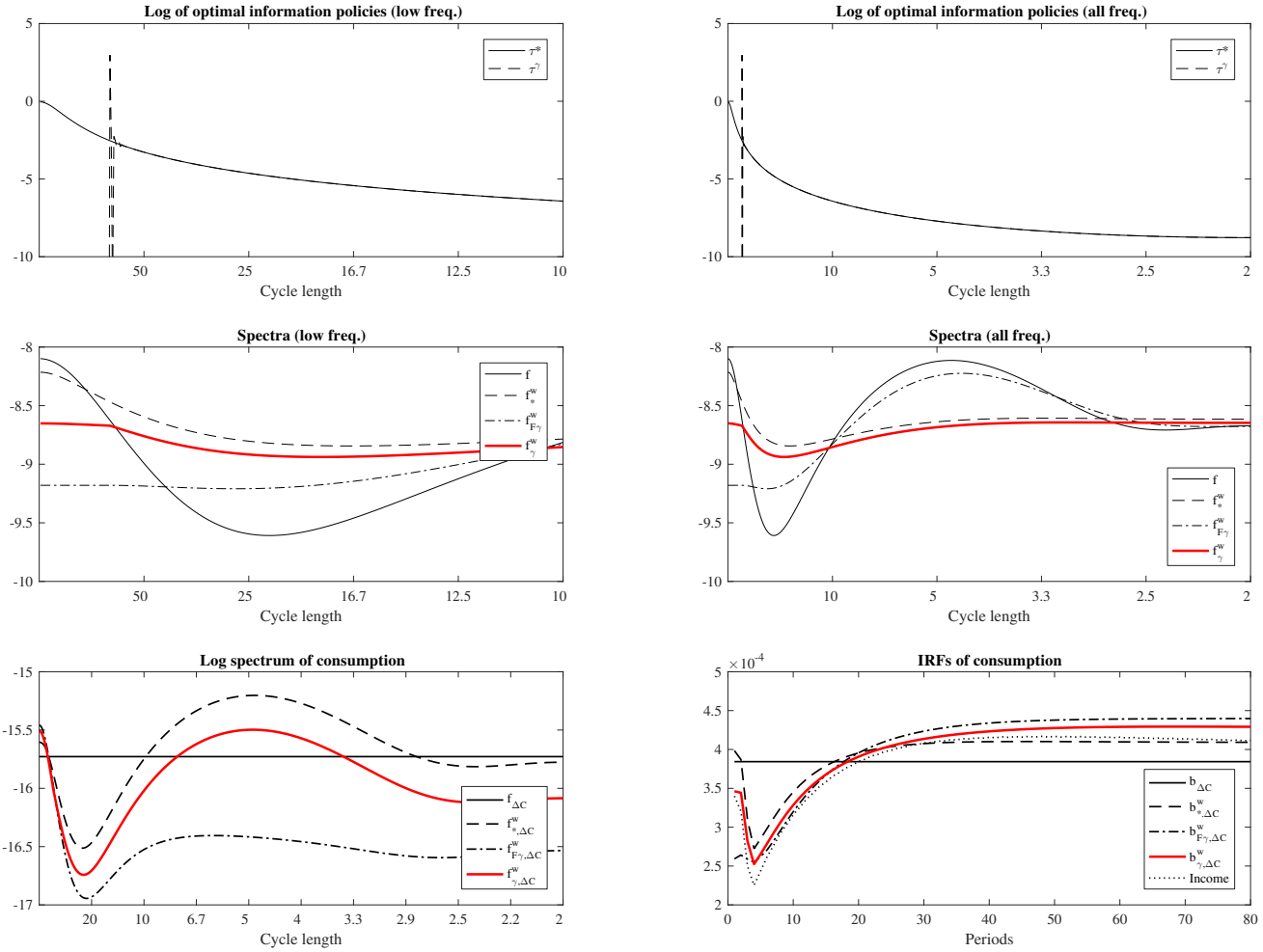
We also examine the correlation matrix of the simulated periodograms. The maximum simulated absolute correlation is 0.0045, while the mean absolute correlation is 0.0008 and the standard deviation is 0.0006. The correlations are therefore uniformly close to zero across the full range of frequencies.

Figure A.1: Variance of spectral estimates with alternative prior



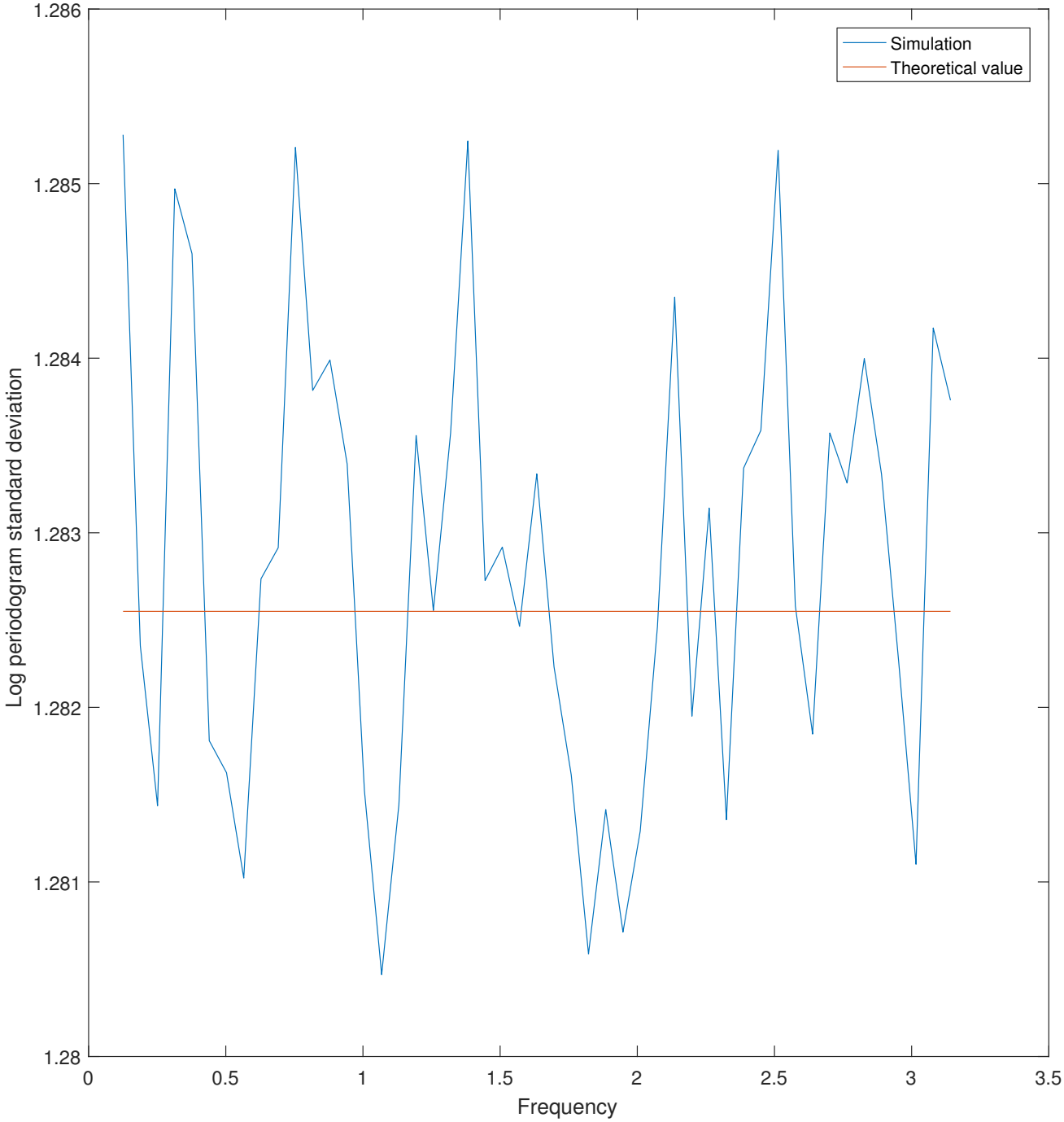
Notes: Variance of estimates of the spectrum across frequencies for the smoothness prior in terms of cycles.

Figure A.2: Effects of no information below a cutoff frequency



Notes: Replicates figure **??**, but instead of assuming information becomes progressively more costly at low frequencies, assumes that information is unavailable below a cutoff frequency (corresponding to 74-year cycles) and that all other frequencies are equally costly to learn about. Note that the first row of panels plots the log of precision, rather than the level.

Figure A.3: Standard deviation of log periodogram across frequencies



Notes: Standard deviation of log periodogram in 100-period simulations of the two-peak spectrum studied in the main text.