Noise in Expectations: Evidence from Analyst Forecasts

Tim de Silva* David Thesmar*†

This version: January 11, 2023

Click here for latest version.

Abstract

Analyst forecasts outperform statistical forecasts in the short run, but underperform in the long run. We develop a framework to decompose these differences in forecasting accuracy into differences in information, forecast bias, and forecast noise. Noise and bias strongly increase with the horizon, while analyst information advantage decreases very fast. Quantitatively, this generates a reversal in the sign of the Coibion and Gorodnichenko (2015) regression coefficient at long horizons, independently of overreaction/underreaction. Finally, we show a parsimonious model with bounded rationality and a noisy cognitive default à la Patton and Timmermann (2010) jointly matches the term structures of noise and bias.

*MIT Sloan

†NBER and CEPR

Corresponding emails: tdesilva@mit.edu; thesmar@mit.edu. We thank Alberto Abadie, Sam Anderson, Anne Beyer, Josefina Cenzón, Harin de Silva, Jack Ely, Cary Frydman, Stefano Giglio, Eric Ghysels, Allen Hu, Simas Kucinskas, Eben Lazarus, Charles Lee, Albert Marcet, Eric So, Frank Schilbach, Andrei Shleifer, Dmitry Taubinsky, two anonymous referees, and seminar participants at various seminars and conferences for helpful comments and suggestions. An earlier version of this paper were circulated with the title "The Term Structure of Subjective Expectations: Evidence and Theory from Analysts Forecasts".

Subjective forecasts can differ from rational expectations in three ways. First, forecasters may have different information sets. Second, forecasts may be biased, meaning their forecast errors are predictable. Finally, forecasts can be noisy, meaning there is variation in forecasts that is unpredictable and unrelated to the realization. While the existing literature studying expectation formation has charactered many ways in which subjective forecasts are biased, this objective of this paper is to quantify expectation noise and analyze its properties. Such expectation noise plays a central role in theoretical models of belief formation (e.g., Sims 2003; Woodford 2003; Afrouzi, Kwon, Landier, Ma, and Thesmar 2021) and is a pervasive feature of forecasting and human decision-making in many domains, such as medicine, finance, hiring, or the justice system (Kahneman, Sibony, and Sunstein 2021).

The approach to quantifying noise in expectations that we develop in this paper is motivated a simple decomposition. To illustrate, denote $F\pi$ as the forecast of a variable π formed by a subjective forecaster, X as the information set of an econometrician, and Z as the set of "soft" information relevant for forecasting π unknown to the econometrician but possibly to the forecaster. Without loss of generality, we can write the subjective forecast as:

$$\underbrace{F\pi}_{\text{subjective forecast}} = \underbrace{E(\pi|X)}_{\text{econometric forecast}} + \underbrace{E(\pi|X,Z) - E(\pi|X)}_{\text{soft information}} + \underbrace{B(X,Z)}_{\text{bias term}} + \underbrace{\eta}_{\text{noise term}}$$

The bias term, $B(X,Z) = E(F\pi - \pi|X,Z)$, is the predictable deviation from rational expectations conditional on available information, which has been studied extensively. The object of interest in this paper is the noise term, $\eta = F\pi - E(F\pi|X,Z)$, which comes on top of the bias, soft information, and the rational component. Such noise can arise in rational models due to noisy information (e.g., Woodford 2003) or irrational models due to cognitive noise (e.g., Afrouzi et al. 2021). Under mild assumptions discussed in Section 2, the previous equation implies the mean squared error (MSE) of the subjective forecast can be written as:

$$\underbrace{MSE}_{\text{subjective MSE}} - \underbrace{MSE^e}_{\text{econometric MSE}} = -\underbrace{E\left[E(\pi|X,Z) - E(\pi|X)\right]^2}_{\text{soft information}} + \underbrace{E\left[B(X,Z)^2\right]}_{\text{bias}} + \underbrace{\mathrm{var}(\eta)}_{\text{noise}}, \tag{1}$$

where MSE^e is the econometric MSE. According to (1), the difference in MSE between subjective forecasters and the econometrician can be decomposed into three components: noise, bias, and the forecasters' informational advantage.

Our goal in this paper to separately estimate the three elements in (1), with a particular focus on expectation noise, and examine what restrictions they place on models of belief formation. In order to do so, we need to compare subjective and econometric forecasts of the same variable. In this paper, we focus on forecasts of corporate earnings for two reasons. First, the earnings forecasts issued by sell-side equity analysts, who are skilled and incentivized forecasters, provide us with a large panel of subjective forecasts across multiple forecasting horizons. Although the approach we develop to quantify expectation noise is applicable to any dataset on subjective forecasts, this large panel of forecasts allows us to impose minimal restrictions on data-generating processes. Secondly, variation in earnings forecasts are intrisically relevant, as movements in cash flow expectations explain a significant fraction, if not most, of asset price movement (Vuolteenaho 2002; De la O and Myers 2021; Bordalo, Gennaioli, La Porta, and Shleifer 2022)

We begin our analysis by comparing the precision of subjective and econometric forecasts, which is the LHS of (1). To form econometric forecasts, we explore several well-known supervised machine learning estimators (e.g., Elastic Net, Random Forest, Gradient-Boosted Trees) with over 200 predictors from past financial statements and stock prices. When we compare these forecasts to those of equity analysts, we uncover a term structure of forecasting accuracy: subjective forecasts dominate econometric forecasts at short forecasting horizons (one, two, three, four and quarters and one year), but underperform our econometric forecasts at long horizons (two, three, and four years).

As illustrated in (1), the relative performance of subjective and econometric forecasts depends on three terms: differences in information, forecast bias, and expectation noise. Thus, our finding that the relative accuracy of these two forecasts varies with the forecasting horizon implies one (or more) of these three terms must be varying across horizons. For example, subjective forecasts may be more accurate at short horizons because of information gathered from discussions with management that is not part of our econometric forecasts. Alternatively, subjective forecasts might be less accurate at long horizons due greater forecast bias from cognitive mistakes (e.g., extrapolation).

We next turn to developing a quantitative framework that allows us to separately identify each of the three terms on the right of (1): soft information, bias, and noise. In contrast to most of the extant literature, our framework places no restrictions on the data generating process. Our approach does rely on three key assumptions about the structure of the belief formation process, which we show are satisfied in commonly-used models of belief formation such as noisy information (Woodford 2003), sticky expectations (Mankiw and Reis 2002), and diagnostic expectations (Bordalo, Coffman, Gennaioli, and Shleifer 2016). These assumptions are necessary for identification because the information set of forecasters is unobservable. Nevertheless, our framework is flexible as it allows for four classic deviations from full-information rational expectations: (i) bias on public information, (ii) soft information observed by the forecaster, (iii) bias on soft information, and (iv) expectation noise.

We next show how to use this framework to estimate these four components separately at different forecasting horizons using an intuitive set of moment conditions. At short forecasting horizons (less than two years), we estimate soft information is an order of magnitude larger than noise and both sources of bias: bias and noise combined are only around 10-20% of soft information. This is consistent with the fact that subjective forecasts dominate econometric forecasts at short horizons. However, when we look at longer forecasting horizons, we uncover an upward-sloping term structure of noise and bias. For example, at a forecasting horizon of three years, expectation noise increases by a factor of three to around 70% of the size of soft information. Bias also increases by a factor of two. Thus, the upward-sloping term structure of noise, coupled with the increase in bias, is what explains the decay in accuracy of subjective forecasts at longer horizons.

To illustrate the quantitative importance of expectation noise in our setting, we explore two of its implications. First, we show that noise makes the Coibion and Gorodnichenko (2015) (CG) regression coefficient a misleading measure of overreaction at longer horizons. Consistently with existing literature, the CG coefficient in our analyst forecast data is positive at short horizons and negative at longer ones. Traditionally, the literature interprets this as evidence of short-term underreaction and long-term overreaction. We show that this interpretation is misleading because noise biases the CG coefficient towards negative values, even when forecasts are underreacting. To do this, we use our estimation to compute the CG coefficient in a counter-factual world where forecasts have no noise. We find that this counter-factual CG coefficient is positive and increasing with horizon, suggesting greater *under*reaction at long horizons. This discrepancy comes from the upward-sloping term structure of noise: at long horizons, noise is larger, making the CG coefficient negative, in spite of underreaction in observed forecasts.

The second implication of the upward-sloping term structure of noise that we explore is its effect on the complementarity between statistical and subjective forecasts. Even at long horizons, our analysis shows that subjective forecasts contain substantial "soft" information. Thus, we can improve forecasting performance by forming "combined" forecasts that include subjective forecasts as an additional predictor when forming statistical forecasts. However, noise weakens this complementarity because it makes it harder to extract the soft information embedded in statistical forecasts: combined forecasts optimally underweight noisy forecasts, thus extracting less soft information. As a result, an upward-sloping term structure of noise implies these combined forecasts should provide little improvement relative to our benchmark econometric forecasts at long-horizons. We show that this is the case in our setting: man is a good complement of machine in the short-run, but not in the long-run.

In the final part of the paper, we examine which models of belief formation can jointly match the term structures of expectation noise and bias that we estimate. We first revisit several canonical models, including models of noisy information (Woodford 2003), bounded rationality (Sims 2003), diagnostic expectations (Bordalo, Gennaioli, Ma, and Shleifer 2020), over-confidence (Daniel, Subrahmanyam, and Hirshleifer 1998), and over-extrapolation (Greenwood and Shleifer 2014; Angeletos, Huo, and Sastry 2020). In their standard formulations, these models all predict downward-sloping term structures for both bias and noise. This is because these models rely on the law of iterated expectations to determine the term structure of forecasts: since forecasters know the true data-generating process, their forecasts shrink towards the unconditional mean at longer horizons. Thus, in standard expectations models, forecasters are more "rational" at longer horizons, at odds with our evidence that bias and noise increase with horizon.

Motivated by the failure of these models, we deviate by exploring a variant of the model from Patton and Timmermann (2010) that has two key components. First, forecasters exhibit a form of bounded rationality in the spirit of Gabaix (2014). Specifically, forecasts are a weighted sum of a cognitive default and the true conditional expectation, with less weight on the former as the latter becomes more accurate. This dependence is assumed, but not micro-founded following Patton and Timmermann (2010). The second key ingredient is the cognitive default that may contain bias and noise. Thus, the model is parsimonious and mostly relies two key horizon-invariant parameters: one that controls the sensitivity the weight on the cognitive default to the precision of rational forecasts, and one that captures the quantity of expectation noise in the cognitive default.

We estimate the parameters of this model by targeting the term structures of bias and noise that we previously characterized. The estimated noise in cognitive defaults is around the same as the variation in the true data generating process, which allows us to match the large average level of noise in the data. Our ability to match the upward *slope* of the bias and noise term structures is driven by a form of bounded rationality: forecasters rely more on their cognitive defaults at longer horizons because the true conditional expectation is less accurate in absolute terms. Although our model has only one parameter that jointly controls the slope of all term structures, we find it matches them quite well. This finding suggests the underlying mechanisms generating bias and noise are linked, echoing the findings of Enke and Graeber (2020).

Given bounded rationality is a crucial ingredient for this model to fit the data, we conclude by exploring how noise varies cross-sectionally with the volatility of the underlying process. Our model makes the qualitative prediction that noise should increase in volatility, which we show is true empirically. However, our model can replicate this relationship reasonably well quantitatively, even though it is estimated entirely using across-horizon rather than cross-sectional moments.

Related literature. Subjective forecast noise is discussed in the large literature on noisy information (e.g., Woodford 2003; Coibion and Gorodnichenko 2015) and in cognitive psychology (e.g., Khaw, Li, and Woodford 2019; Woodford 2020; Enke and Graeber 2020; Kahneman et al. 2021; Afrouzi et al. 2021). Our contribution to this literature is twofold: (i) our evidence on the size and term structure of noise (in contrast to bias) using analyst forecast data and (ii) our methodology that places no restrictions on the data-generating process. Our methodology is similar in spirit to Satopää, Salikhov, Tetlock, and Mellers (2020), who perform a bias-information-noise ("BIN") decomposition and find a consistent property of good subjective forecasters is noise reduction, and is complementary to the approach developed by Juodis and Kucinskas (2019), which exploits the factor-structure in expectations implied by many models of belief-formation. More broadly, our approach is related to Bianchi, Ludvigson, and Ma (2020) and Nagel (2021), who discuss how supervised learning is useful for studying subjective expectations data.

Our work also connects to the extant empirical literature on expectations formation. This literature generally focuses on estimating forecaster bias (e.g., Manski 2017) and forecaster information *dis*advantage (e.g., Coibion and Gorodnichenko 2015). In contrast, we measure two additional components: subjective forecasters' information advantage and noise. We further document the term structure of these components and explore a modeling assumption – reliance on noisy default – that allows us to fit the data.

Our finding of an upward-sloping term structure of noise is, to our knowledge, novel. Patton and Timmermann (2010) document that disagreement in macro forecasts increases with horizon, which is consistent with this observation. Additionally, this upward-sloping term structure of noise provides support for theories of discounting based by horizon-increasing misperception rather than a fundamental time preference (Gabaix and Laibson 2017), which have received experimental support (Gershman and Bhui 2020). The upward-sloping term-structure of bias is consistent with emerging evidence in asset prices and expectations data (Giglio and Kelly 2018; Bordalo, Gennaioli, La Porta, and Shleifer 2019; D'Arienzo 2020; Angeletos et al. 2020; Afrouzi et al. 2021). More precisely, this literature documents more overreaction at long horizons, which our decomposition of the Coibion and Gorodnichenko (2015) coefficient shows is likely influenced by the presence of expectation noise. Closely related evidence is presented in Dessaint, Foucault, and Frésard (2020), who show that long-term forecasts are less predictive of future earnings realization. This is consistent with long-term forecasts being either more biased, or noisier, or less informed. Our decomposition clarifies this without making assumptions about the true DGP.

Because we estimate statistical forecasts, our paper also engages with the recent literature applying supervised machine learning in economics and finance (see Mullainathan and Spiess 2017, for a review). To perform our decomposition, we study the predictability of corporate earnings at various horizons using firm-level observables and standard ML estimators. So (2013) proposes a parsimonious regression model to forecast EPS, upon which multiple recent papers have expanded by applying ML techniques (see Ball and Ghysels 2018; van Binsbergen, Han, and Lopez-Lira 2020; Hansen and Thimsen 2020; Cao and You 2020). Like these papers, and like papers implementing a similar exercise on equity returns directly (e.g., Gu, Kelly, and Xiu 2018; Kozak, Nagel, and Santosh 2020; Bryzgalova, Huang, and Julliard 2020), we find that there are gains to using supervised ML techniques over non-regularized estimators. Another outcome of our analysis is that tree-based forecasts marginally dominate penalized methods – this is also consistent with the literature on EPS forecasting.

Finally, our paper is related to the extensive literature on analyst forecasts (see Kothari, So, and Verdi 2016, for a review). Our findings that analyst forecasts are more accurate at a horizon of less than a year is broadly consistent with this literature (e.g., Brown and Rozeff 1978; Bradshaw, Drake, Myers, and Myers 2012), and our estimate of a large soft information corroborates the survey evidence in Brown, Call, Clement, and Sharp (2015). Our proposed model features a form of bounded rationality, consistent with evidence of attention constraints shaping analyst forecast behavior by affecting effort allocation (Harford, Jiang, Wang, and Xie 2019) and inducing social learning (Kumar, Rantala, and Xu 2021).

1 The Term Structure of Forecasting Accuracy

1.1 Data Description

The data used in this paper comes from three sources: I/B/E/S, Compustat, and CRSP. We start by collecting the reported fiscal-year-end (FY) earnings-per-share (EPS) and their respective announcement dates from the I/B/E/S actuals file for all US firms with announcements between 1989 and 2021. For each FY denoted as t, we collect all analyst EPS forecasts from the I/B/E/S detailed file issued within 45 calendar days of the release of the FY annual report[1]. We focus on this 45-day period to make sure our subjective forecasts are not stale and are taken with similar information sets across analysts (as in Bouchaud, Krüger, Landier, and Thesmar 2019). We use all available quarterly forecasts and all annual forecasts excluding the five-year-ahead forecasts (due to a lack of observations). When analysts issue multiple forecasts, we keep only their earliest

[1]Our results are robust to using a 30-day instead of 45-day window.

forecasts.

We denote forecasting horizons by h, where $h \in \{1,2,3,4\}$ denotes annual forecasts and $h \in \{0.25, 0.5, 0.75, 1^*\}$ denotes quarterly forecasts. For each forecasting horizon, we collect the corresponding realization of EPS from the I/B/E/S actuals file. We then normalize both forecasts and realizations by the stock price from CRSP on the day of the fiscal year-end.[2] We denote the realizations of this earnings-to-price ratio for firm i at time $t+h$ as $\pi_{it+h} = \frac{EPS_{it+h}}{P_{it}}$ and corresponding forecasts by $F_t^j \pi_{it+h} = \frac{F_t^j EPS_{it+h}}{P_{it}}$, where j indexes analysts and $F_t^j EPS_{it+h}$ is the forecast at horizon h by analyst j. So we normalize both forecasts and eventual realizations, at all horizons, by the same price P_{it}.

Next, we collect a large set of financial ratios from the Financial Ratios provided by WRDS. We also collect several variables from Compustat, CRSP, and IBES. We denote the set of these variables as X_{it}, all of which are listed in Table A1. Each of these variables is calculated using information available upon the release of the fiscal year-end in year t for firm i. Finally, we impose several sample filters: we delete all observations for securities that are not ordinary equity securities (CRSP share codes 10 and 11), winsorize forecasts of EPS and EPS at 10 times their interquartile range to eliminate outliers, and drop a small number of observations in which forecast errors are extremely large, which are likely data errors.

In Table 1, we show several summary statistics on the set of firm-year-analyst observations in our sample. Panel A shows the average forecast error across the four quarterly and four annual horizons we examine. Looking at the mean forecast error across horizons, we already see evidence of an upward-sloping term structure of forecast bias: forecasts exceed the realization on average and this difference increases with the forecasting horizon. In Panel B, we show summary statistics on the set of firm-years that are in our sample at each forecasting horizon. For all quarterly, one-year, and two-year-ahead forecasts, we observe around 4-5 distinct analyst forecasts per firm. Coverage drops after that, both in terms of total number of forecasts, and forecasts per firm. In terms of size, the firms at these different horizons appear relatively similar. However, at the two longest forecast horizons, three- and four-year-ahead forecasts, we have distinct firms and forecasts per firm. This is because there are much fewer forecasts available in I/B/E/S at these longer horizons. As expected, in terms of size the firms for which we have forecasts at longer horizons tend to be larger.

[2] We work with earnings-to-price ratios instead of EPS levels because this variable has substantially fewer outliers. Our results are robust to using EPS levels.

Table 1. Summary Statistics

This table shows summary statistics on our final sample that we construct as described in Section 1.1. In Panel A, we show summary statistics for forecast errors at the firm-year-analyst-level for our different forecasting horizons. Panel B shows summary statistics at the firm-year level of the number of distinct analysts, N_{it}, and the total assets in $ millions.

Panel A: Analyst Forecasts

	Count	Mean	SD	10%	25%	50%	75%	90%
$\pi_{it+h}^{h=0.25} - F_t^j \pi_{it+h}^{h=0.25}$	358,299	0	0.005	-0.004	-0.001	0	0.002	0.005
$\pi_{it+h}^{h=0.5} - F_t^j \pi_{it+h}^{h=0.5}$	311,253	-0	0.008	-0.007	-0.002	0	0.002	0.006
$\pi_{it+h}^{h=0.75} - F_t^j \pi_{it+h}^{h=0.75}$	305,763	-0.001	0.010	-0.010	-0.003	0	0.002	0.006
$\pi_{it+h}^{h=1^*} - F_t^j \pi_{it+h}^{h=1^*}$	305,844	-0.002	0.012	-0.014	-0.004	-0	0.002	0.006
$\pi_{it+h}^{h=1} - F_t^j \pi_{it+h}^{h=1}$	388,895	-0.004	0.031	-0.033	-0.009	0	0.005	0.017
$\pi_{it+h}^{h=2} - F_t^j \pi_{it+h}^{h=2}$	308,682	-0.012	0.052	-0.062	-0.022	-0.003	0.006	0.023
$\pi_{it+h}^{h=3} - F_t^j \pi_{it+h}^{h=3}$	51,722	-0.018	0.072	-0.087	-0.035	-0.006	0.007	0.032
$\pi_{it+h}^{h=4} - F_t^j \pi_{it+h}^{h=4}$	12,816	-0.037	0.110	-0.142	-0.059	-0.013	0.006	0.033

Panel B: Firm-Level Variables

	Count	Mean	SD	10%	25%	50%	75%	90%
$N_{it}^{h=0.25}$	75,970	4.790	4.309	1	2	3	6	10
Total Assets$_{it}^{h=0.25}$	75,970	9.605	68.532	0.077	0.217	0.834	3.305	12.880
$N_{it}^{h=0.5}$	70,250	4.489	4.166	1	2	3	6	10
Total Assets$_{it}^{h=0.5}$	70,250	10.151	70.751	0.080	0.229	0.879	3.494	13.744
$N_{it}^{h=0.75}$	68,858	4.489	4.154	1	2	3	6	10
Total Assets$_{it}^{h=0.75}$	68,858	10.281	71.711	0.081	0.232	0.891	3.546	13.882
$N_{it}^{h=1^*}$	66,983	4.610	4.242	1	2	3	6	10
Total Assets$_{it}^{h=1^*}$	66,983	10.372	72.149	0.082	0.236	0.917	3.671	14.190
$N_{it}^{h=1}$	75,782	5.189	4.863	1	2	4	7	12
Total Assets$_{it}^{h=1}$	75,782	9.478	68.437	0.068	0.202	0.794	3.200	12.505
$N_{it}^{h=2}$	64,241	4.839	4.421	1	2	3	6	11
Total Assets$_{it}^{h=2}$	64,241	10.134	69.989	0.081	0.240	0.924	3.617	13.838
$N_{it}^{h=3}$	20,660	2.518	2.173	1	1	2	3	5
Total Assets$_{it}^{h=3}$	20,660	20.607	109.421	0.163	0.614	2.470	9.321	32.884
$N_{it}^{h=4}$	7,936	1.630	1.302	1	1	1	2	3
Total Assets$_{it}^{h=4}$	7,936	23.471	110.206	0.131	0.533	2.964	13.999	44.208

1.2 Forecast Formation

First, we calculate consensus analysts forecasts, which we denote by $F_t \pi_{it+h}$. We calculate consensus forecasts by taking an equally-weighted average of the analyst forecasts we have in our sample for each firm-year in Table 1.

Next, we turn to the formation of our statistical (or "econometric") forecasts, which we denote by $F_t^e \pi_{it+h}$. Given a set of public information X_{it}, our goal is to approximate the conditional expectation function, $E(\pi_{it+h}|X_{it})$, as accurately as possible. As is well-known, $E(\pi_{it+h}|X_{it})$ is the solution to the problem of minimizing mean squared error across all possible (measurable) functions of X_{it}:

$$E_t(\pi_{it+h}|X_{it}) = \underset{h(X_{it})}{\arg\min} E\left[(\pi_{it+h} - h(X_{it}))^2\right] \tag{2}$$

In practice, solving (2) is infeasible because it requires searching over an infinite dimensional function space. To gain tractability, we leverage supervised machine learning approaches that restrict $h(X_{it})$ to be within a particular class of functions, such as linear functions, and use different forms of regularization developed in supervised machine learning to address the high dimensionality of the set of variables in X_{it}.

The first step is to decide what variables constitute X_{it}. To form X_{it}, we use all the variables listed in Table A1, which consist of a large set of commonly-used financial ratios, industry indicator variables, and past stock price information. We use these variables for fiscal year t and from the prior two annual (or quarterly) reports to capture potential lead-lag relationships, resulting in a set of over 200 predictor variables. Our logic for choosing these variables is not that we think they represent the exhaustive set of information relevant for forecasting earnings at the firm-level. Instead, we view these variables as a large set of variables that are easily observable and likely used by analysts, against which we will compare predictive power. Importantly, since the econometrician is forecasting π_{it+h} with X_{it}, it needs to wait until after the release of the fiscal year report in year t before forming its forecasts. This econometric forecast can therefore be thought of as being issued at the same calendar time as the analyst forecast we collect from I/B/E/S.

To solve the sample counterpart to (2), we next need to define the training sample. To avoid look-ahead bias, we use rolling windows of 5 years to train the various statistical forecasting models, with the exception of four-year-ahead forecasts.[3] This period is chosen to maximize the size of

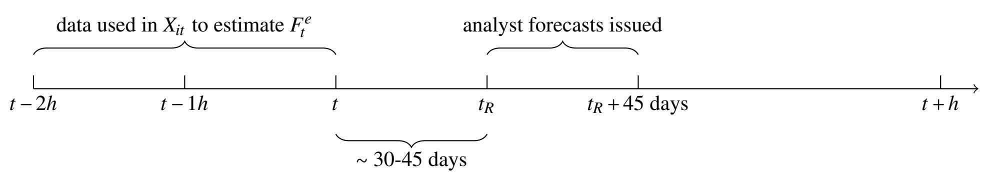
[3] Due to the fewer number of observations for $h = 4$, we use an expanding window in which past data accumulates and is never dropped from training in future years.

the training set subject to computational constraints.[4] Additionally, by choosing a rolling window, we implicitly allow for low-frequency changes in the data generating process for earnings over time. More precisely, for each date t, we use all firm-year observations (i,s) in years $s \in \{t-4, t-3, ..., t\}$ to forecast at π_{it+h} for our eight forecasting horizons, $h \in \{0.25, 0.5, ..., 1, ..., 4\}$. Given forecasting variables X_{is}, our estimation consists of finding the function that has the minimum MSE in explaining in-sample future $\pi_{i,s+h}$, using various regularization techniques described below. We refer to this estimated function as our *econometric forecast*, and denote it by $F_t^e \pi_{it+h}$.

Figure 1 provides a timeline to clarify the relative timing of our data collection, analyst forecasts, and econometric forecasts. Consider a firm with an annual fiscal year-end at t, which releases the annual report at t_R (typically 30-45 calendar days after t). The data we use to form econometric forecasts is realized at t, $t-1h$, and $t-2h$. Thus, these econometric forecasts are "current" at the time of t_R, when this information is made public. We then collect analyst forecasts over the 45 days following t_R so that these forecasts are not stale and are made with access to the information set X_{it}, upon which our econometric forecasts are based.

Figure 1. **Timeline of Forecast Formation**

This figure provides a timeline of the formation of our forecasts of π_{it+h}. t denotes the fiscal-year-end, while t_R denotes the release of the fiscal-year-end report. $t+h$ denotes the time at which the realization, π_{it+h}, is realized.



Finally, we describe the supervised learning techniques we use to estimate the forecasting function that solves (2), F_t^e. We choose to depart from using OLS because our goal is approximation of a conditional expectation functions, so we'd like to impose minimal function form restrictions. If X_{it} was low-dimensional, we could in principle use OLS, but since X_{it} is not low-dimensional, OLS is inconsistent and unstable due to its tendency to overfit. Thus, we turn to supervised learning techniques, which restrict the spaces of possible functions in (2) to be tractable yet flexible, while simultaneously minimizing the risk of over-fitting using various forms of regularization. The following is a brief presentation of our approach; we refer the reader to Appendix C for more discussion on the theoretical properties of these estimators and our implementation.

[4]To check that our conclusions are not sensitive to the choice of a 5-year window, we ran one of our estimators (Gradient-Boosted Trees) using a growing window, where all past data is used (this is too computationally challenging for our penalized linear estimators, as we include many interactions in those estimations). We found the MSE of the econometric forecast reduced only by 2.0%, and the econometrician + analyst forecast MSE increased by 0.6%.

Random walk. As a benchmark, we consider a random walk forecast where $F_t^e \pi_{it+h} = EPS_{it} \ \forall h$. Although this will be our worst performing forecast, it is benchmark commonly used in literature on analyst forecasts (e.g., Bradshaw et al. 2012).

Elastic net. The first supervised learning estimator we consider is Elastic Net. This estimator, which is a penalized linear estimator, is defined by the solution to the same objective function that OLS solves (minimizing in-sample mean squared error), but with an additional penalty term on the size of the coefficients, where the size of the penalty is chosen via cross-validation on the training set (detailed in Appendix C). Intuitively, cross-validation consists of breaking up the training sample into smaller datasets, fitting models on these smaller datasets, and examining which penalty value generates the best performance on the other parts of the training set. Importantly, cross-validation is done entirely using the training set to avoid introducing any look-ahead bias.[5]

Random forest. The second estimator we consider is Random Forest (RF), which is a non-parametric tree-based method. The building block of tree-based estimators are regression trees, which are nonparametric regression estimators designed to capture arbitrary non-linearities among the variables in X_{it}. Used alone, regression trees have a tendency to over-fit, which led to the development various "ensemble" methods that introduce forms of regularization. RF is a particular ensemble method constructed based on the intuition of bootstrapping. On each bootstrapped sample, a regression tree is grown. After doing this multiple times, final predictions from the RF are calculated by averaging predictions across the multiple regression trees. Averaging across many trees, which have different structures due to the randomness in the subset of predictor variables chosen, is the regularization in this method that limits over-fitting and reduces prediction variance. Similar to the penalized linear estimators, the parameters that govern the shape of the regression trees can be chosen using cross-validation on the training set.

Gradient-boosted trees. Gradient-boosted trees (GBT) are a second tree-based ensemble method based on the same core idea as RF: grow a large number of uncorrelated trees and then average their predictions. GBT start by fitting a very shallow tree, meaning only a small number of variables are used. This shallow tree is likely has terrible in-sample fit. To improve its fit, a second shallow tree is fit on the residuals calculated from the first tree. Predicted values are then formed by using a weighted average of the predicted values from the two trees. This procedure is repeated many times, after which the predicted value will be a weighted-average of the predicted value from all the shallow trees. The sequential growing of trees on residuals from the previous trees makes the trees less correlated, which is why averaging over trees limits over-fitting. As with the other

[5]We have explored other penalized linear estimators, such as Lasso, Ridge, Post-Lasso (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins 2016), and Iterative-Lasso (Belloni, Chernozhukov, and Hansen 2011), all of which give nearly identical results. We choose to omit them for brevity.

methods, the parameters governing the size of these shallow trees and the weights in the weighted average are chosen using cross-validation on the training set.

1.3 Forecasting Results

We now compare the relative accuracy of our analyst and econometric forecasts discussed in Section 1.2. We present the out-of-sample mean squared errors of all of our forecasts across the entire sample, normalized by the mean squared realization of realized profits π_{it+h}. This normalization can be interpreted as the allocative efficiency loss compared to a perfect foresight optimizer. We discuss this model formally in Appendix D.

Table 2. The Term Structure of Forecasting Accuracy: Analyst vs. Econometrician

This table contains the mean squared error of analyst forecasts in the first column, denoted MSE_h^a, and of our econometric forecasts, denoted MSE_h^e, across different forecasting horizons when forecasting earnings yields, π_{it+h}. The numbers reported in the table are normalized by the mean realization of π_{it+h}^2 at each horizon. In parentheses, we report the Diebold-Marino test-statistics for testing the relative accuracy of the two forecasts under a squared loss function, where the asymptotic variance is calculated by performing a bootstrap at the year level with 1,000 iterations.

	MSE_h^a	MSE_h^e			
Horizon: h	Analyst	Random Walk	Elastic Net	Random Forest	Boosted Trees
1 Quarters	4.6%	26.1%	20.81%	15.52%	17.25%
		(24.34)	(20.45)	(17.73)	(18.45)
2 Quarters	8.43%	30.0%	19.61%	15.26%	16.91%
		(21.45)	(17.89)	(14.56)	(15.06)
3 Quarters	13.05%	33.87%	22.05%	17.87%	19.28%
		(17.93)	(17.86)	(12.04)	(13.46)
4 Quarters	18.71%	24.92%	25.34%	21.55%	22.45%
		(8.86)	(11.92)	(5.93)	(6.73)
1 Years	9.9%	18.67%	18.63%	15.78%	16.52%
		(12.5)	(15.98)	(12.51)	(12.33)
2 Years	29.19%	34.27%	30.21%	27.05%	29.28%
		(4.07)	(1.04)	(-2.13)	(0.08)
3 Years	33.32%	35.11%	29.11%	26.32%	27.96%
		(1.78)	(-5.12)	(-8.97)	(-6.14)
4 Years	46.41%	37.26%	27.86%	25.76%	29.3%
		(-6.56)	(-14.64)	(-15.98)	(-13.39)

Table 2 shows the normalized MSE for our eight forecasting horizons.[6] Focusing first on analyst forecasts, the first column shows that they are very accurate at short horizons. For example, for one-

[6]The decline in MSE levels between four quarter-ahead and one year-ahead forecasts comes from the fact that

quarter forecasts, analyst forecasts only generate a 4% loss in utility relative to a perfect foresight optimizer. This is consistent with the fact that near-term analyst forecasts are heavily influenced by discussions with management, and hence are well-informed. As the forecasting horizon increases, Table 2 shows the relative accuracy of analyst forecasts monotonically declines. At a four-year forecast horizon, we see that the normalized MSE is ten-times as large as for one-quarter forecasts.

The remaining four columns of Table 2 show the results from our econometric forecasts, which are formed with four different methods, in addition to Diebold-Marino test statistics for the relative accuracy of analyst to econometric forecasts under a squared loss function. The first takeaway from these columns is that there are gains to using both more information and less parametric supervised learning methods to forecast earnings. Comparing the second and third columns, we see Elastic Net outperforms Random Walk forecasts by a larger margin at short horizons, and a smaller margin at longer horizons. The final two columns show tree-based methods perform even better at all horizons, generating around a 10-20% improvement relative to Elastic Net. These findings are consistent with existing literature on forecasting firm-level EPS (Ball and Ghysels 2018; van Binsbergen et al. 2020; Hansen and Thimsen 2020; Cao and You 2020), which finds that more sophisticated estimators improve the quality of short-term predictions. We find here this to be the case, not only at short horizons, but also at longer horizons.

The second and more important takeaway from these columns is that analyst forecasts dominate all econometric forecasts at forecast horizons of less than one-year (the DM statistics reject the null at 1% critical values). This is especially true at the one-quarter and two-quarter horizons, where the difference is extremely large: our best econometric forecast (Random Forest) generates a utility loss of around 15% relative to perfect foresight, which is around 2-3 bigger than that of analyst forecasts. However, at longer horizons of two, three, and four years, we find our best econometric forecast (Random Forest) outperforms analyst forecasts. At three-year and four-year horizons, this difference is substantial: Random Forest generates a gain in MSE relative to analyst forecasts of around 10-20pp (33-45%).

Finally, the evidence in Table 2 speaks to a growing literature that compares the relative accuracy of supervised learning estimators designed to perform well under (approximate) sparsity conditions (i.e. variants of Lasso and Elastic Net) to non-sparse or less-parametric estimators (e.g., Ridge regression, tree-based estimators). Our finding that tree-based methods outperform Elastic Net suggests the true data generating process may not be very sparse. This is consistent with from recent evidence in empirical asset pricing that sparse approximations to stochastic discount factors

quarterly forecasts are for each individual quarter. The one-year ahead MSE should instead be compared to the average of the four quarterly MSEs, to which it is similar.

(Gu et al. 2018; Kozak et al. 2020; Bryzgalova et al. 2020) perform poorly, but contrasts with the strong performance of sparse estimators for forecasting macro aggregates (Bianchi et al. 2020).

In sum, this section documents a *term structure of forecasting accuracy*: subjective forecasts are more accurate than statistical forecasts at short horizons, but their accuracy decays at longer horizons. In principle, this reduction in relative subjective forecast accuracy could occur for two reasons. First, subjective forecasters could have access to less soft information at longer horizons. For example, analysts may receive strong signals from discussions with management about a firm's near-term prospects or use high-frequency data sources (Dessaint et al. 2020), which might be less valuable (in terms of forecasting MSE) at longer horizons. Secondly, analysts may issue more biased or noisy forecasts at longer horizons, possibly driven by weaker forecast incentives, a greater tendency to engage in cognitive mistakes (e.g., extrapolation), or a greater cost of processing public information. Quantifying these competing explanations requires a quantitative framework, which we develop and estimate in the next section.

2 Decomposing the Term Structure of Subjective Forecasts

2.1 Framework

2.1.1 MSE Decomposition

We denote $\pi_i = \pi_{it+h}$ as the state variable we seek to forecast (the Data Generating Process, or DGP), which is the earnings-to-price ratio of firm i realized at $t+h$. Throughout this section, we suppress indices t and h to lighten notation, as our analysis imposes no restrictions across t or h. We consider two information sets that can serve to forecast π_i.[7] First, we denote public information observed by the econometrician as X_i, which we assume includes a constant. In our empirical analysis, this corresponds to the set of variables in Table A1. Secondly, we denote Z_i as the second information set that is unobserved to the econometrician and *possibly* observed by analysts, which may intersect with, or contain, X_i. Again, these two information sets do depend on t and h, but we omit these indices for clarity.

Our first result is a decomposition of the DGP that holds without loss of generality.

[7] Throughout we use the term "information set" to informally refer to a sub-σ-algebra on the probability space over which π_i is defined.

Lemma 1. *For any information structure X_i and Z_i, we can decompose the DGP π_i as:*

$$\pi_i = x_i + z_i + \varepsilon_i,$$

where

- $x_i \equiv E(\pi_i \mid X_i)$ *is the component observable to the econometrician;*
- $z_i \equiv E(\pi_i \mid X_i, Z_i) - E(\pi_i \mid X_i)$ *is the soft information component, for which* $E(z_i|x_i) = 0$*;*
- $\varepsilon_i \equiv \pi_i - E(\pi_i \mid X_i, Z_i)$ *is the unpredictable residual, for which* $E(\varepsilon_i|x_i, z_i) = 0$.

All derivations are provided in Appendix A. Lemma 1 states that, without loss of generality, we can decompose π_{it+h} into a part that depends on public information, x_i, an orthogonal part that depends on public *and* non-public information, z_i, and an innovation relative to both information sets, ε_i. If there is no non-public information (i.e. $Z_i \subseteq X_i$), then $z_i = 0$. For this reason, we refer z_i as soft information: it captures the extent to which rational forecasts of π_i change when conditioning on Z_i in addition to X_i.

Forecasts of the DGP π_i are made by forecasters that are indexed by j, which we denote by $F_j\pi_i$. Using the terms defined in Lemma 1, we similarly decompose subjective forecasts, without loss of generality.

Lemma 2. *For any information structure X_i and Z_i, we can decompose the forecast $F_j\pi_i$ as:*

$$F_j\pi_i = x_i + z_i + b_{ij} + \eta_{ij}, \tag{3}$$

where

- $b_{ij} = E\left(F_j\pi_i - \pi_i \mid X_i, Z_i\right)$ *is the analyst bias;*
- $\eta_{ij} = F_j\pi_i - E\left(F_j\pi_i \mid X_i, Z_i\right)$ *is the analyst noise, for which* $E\left(\eta_{ij}|x_i, z_i\right) = 0$.

Equation (3) is the equation described in the introduction, which breaks subjective expectations into three parts. First, the rational expectation given X_i and Z_i, which is $x_i + z_i$. The second term captures forecaster bias, b_{ij}, which represents forecast errors that are predictable based on both information sets. Bias may arise for many reasons, such as behavioral expectation errors, or incentives structures that change forecasters' objectives away form minimizing forecast MSE (e.g., Chen and Jiang 2006). We take no stand on the source of either bias. The final term, η_{ij}, is the noise term. Note that this is just a decomposition that holds without any restrictions: a structural interpretation of each of these terms requires a choice of Z_i, which we specify below.

To understand what expectation noise captures, it is helpful to distinguish Z_i from the information set used by the forecaster j to make her forecast, which we denote as Z_{ij}. Using this notation, noise can be broken down into two parts:

$$\eta_{ij} = \underbrace{\left[E\left(F_j\pi_i \mid X_i, Z_{ij}\right) - E\left(F_j\pi_i \mid X_i, Z_i\right)\right]}_{\text{observation noise}} + \underbrace{\left[F^j\pi_i - E\left(F^j\pi_i \mid X_i, Z_{ij}\right)\right]}_{\text{“Kahneman” noise}},$$

The first part of noise comes from the fact that the forecaster may not have the “true” information set, meaning Z_i differs from Z_{ij}. For example, consider noisy information models (e.g., Woodford 2003) where each forecaster receives a signal that is a noisy version of Z_i, but is rational. Then, the first term will be non-zero and there will be noise (see Example 2 below). The second source of noise is captured by the second bracketed term: variation in forecasts that cannot be explained by forecasters' information sets. In most models of expectation formation used in economics and finance, this term is zero – two forecasters with the same information sets make the same forecasts. However, in general this need not be true, as illustrated by the numerous examples in Kahneman et al. (2021). One possible micro-foundation for such noise is the large evidence in cognitive psychology of individuals' noisy retrieval and storage of information (which has been recently analyzed in Khaw et al. 2019; Woodford 2020; Enke and Graeber 2020; Afrouzi et al. 2021).[8]

We next turn to deriving our MSE decomposition. We perform this decomposition using consensus forecasts since these are available for all firms, but use individual analyst forecasts for estimation (more details on this below). Letting J_i denote the number of analysts issuing forecasts on firm i, we define the consensus forecast as the mean forecast across all forecasters:

$$F\pi_i = \frac{1}{J_i}\sum_{j=1}^{J_i} F_j\pi_i = x_i + z_i + \underbrace{\frac{1}{J_i}\sum_{j=1}^{J_i} b_{ij}}_{\equiv b_i} + \underbrace{\frac{1}{J_i}\sum_{j=1}^{J_i} \eta_{ij}}_{\equiv \eta_i}.$$

Here b_i and η_i represent the bias and noise terms in the consensus forecasts, respectively. We make the rather weak assumption that $J_i \in (X_i, Z_i)$, which is true in our empirical application where it is

[8]Any elicitation noise and classical measurement error would also generate this second type of noise. However, we do not emphasize this interpretation because there is little reason to expect these to be large in our setting. Moreover, even if they were, we don't see a reason to expect them to vary over the forecasting horizon when we use newly updated forecasts.

part of X_i.

We now provide the decomposition of the MSE of subjective forecasts shown in (1) in the Introduction. Define $MSE^a = E\left[(F\pi_i - \pi_i)^2\right]$ be the MSE of consensus forecasts and $MSE^e = E\left[(x_i - \pi_i)^2\right]$ be the MSE of econometric forecast. The following lemma states the result.

Lemma 3 (MSE decomposition)**.** *Assume that the DGP innovation is uncorrelated with expectation noise: $E\left(\varepsilon_i \eta_{ij}\right) = 0$. Then, the difference between the MSE of consensus and econometric forecasts is:*

$$MSE^a - MSE^e = -E\left(z_i^2\right) + E\left(b_i^2\right) + var\left(\eta_i\right).$$

This decomposition formalizes the discussion of the end of Section 1.3. Analyst can outperform statistical forecasts if they have soft information (large $E\left(z_i^2\right)$), have low bias (small $E\left(b_i^2\right)$) and low noise (small $\mathrm{var}\left(\eta_i\right)$). Since the accuracy of consensus forecasts deteriorates at longer horizons (Section 1.3), longer-term forecasts must have less soft information, more bias, or more noise.

Our goal is to estimate these three components. The challenge in doing so is that Z_i (and Z_{ij}) is not observed. In order to make progress on identification, we need to place more structure on the data. Our approach in this paper is to avoid making assumptions on the DGP for π_i, but instead make assumptions on the structure of forecasts. We next turn to discussing these assumptions.

2.1.2 Structural Assumptions

The structural assumptions on the data generating process for forecasts that we work with for the remainder of the paper are stated Assumption 1.

Assumption 1. *Subjective forecasts, $F_j\pi_i$, satisfy the following conditions:*

1. *Forecaster bias on non-public information is proportional to soft information z_i:*

$$b_{ij} - E\left(F\pi_i - x_i \mid X_i\right) = (\alpha - 1)z_i, \tag{4}$$

2. *Expectation noise is conditionally uncorrelated with the DGP innovation:*

$$E\left(\varepsilon_i \cdot \eta_{ij} \mid X_i, Z_i\right) = 0. \tag{5}$$

3. *Expectation noise is conditionally uncorrelated across forecasters:*

$$E\left(\eta_{ij}\cdot\eta_{ik}\,|\,X_i,Z_i\right)=0,\quad \forall\, j\neq k. \tag{6}$$

4. *The square of expectation noise is mean independent of the number of analysts:*

$$E\left(\eta_{ij}^2\,|\,J_i\right)=E\left(\eta_{ij}^2\right)=var\left(\eta_{ij}\right). \tag{7}$$

Let us discuss these assumptions. Equation (4) is our most significant structural assumption. It embeds three restrictions. First, it assumes that bias on public and soft information are separable, which we view as a natural starting point given existing models of expectations formation with multiple information sources (e.g., Chen and Jiang 2006; Maćkowiak and Wiederholt 2009; Kacperczyk, Van Nieuwerburgh, and Veldkamp 2016). This is because the left-hand side of (4) is the residual bias that remains after projecting out the bias on public information, X_i. Second, it requires that the residual bias, which is on soft information, is linear in the true quantity of information, z_i, where $\alpha=1$ corresponds to the case of no bias. Such linearity is necessary for identification. This restriction could be viewed as a first-order Taylor approximation around the mean of z_i. Finally, (4) requires bias to be constant across forecasters. This assumption is in line with the existing literature: heterogeneity in biases is hard to estimate, especially when these biases concern unobserved information.

The three remaining conditions in Assumption 1, (5)-(7), place restrictions on the noise term. Equation (5) requires forecasting noise to have no direct effect on realizations – it is already necessary to obtain the main decomposition in Lemma 3. It would fail, for example, in a model where investors' noise about price forecasts would itself affect aggregate demand and thus equilibrium prices. Equation (6) imposes that the forecaster noise term, η_{ij}, is uncorrelated across forecasters, which is consistent with the two broad interpretations of noise discussed above: noisy information (e.g., Woodford 2003) and cognitive noise (e.g., Kahneman et al. 2021). Equation (7) ensures that the variance of noise is uncorrelated with the number of analysts following a firm. It would fail, for instance, if more complex firms are followed by more analyst with noisier expectations.

Assumption 1 is not generically satisfied, but we now show that it holds in the several existing models of expectations formation, provided Z_i is properly defined. As a result, we view Assumption 1 as a good starting point for decomposing the term structure of forecasts. Even with these restrictions, our framework is quite rich: we allow for unrestricted bias on public information, unobserved private information, bias on unobserved information, and noise, all with no restrictions on the DGP for EPS or across forecasting horizons. In contrast, many papers in the literature focus on AR1 processes with no unobserved information.

Example 1: full-information rational expectations. Full-information rational expectations are defined by:

$$F_j \pi_i = E\left(\pi_i \mid Z_{ij}\right).$$

If analysts have full information, X_i is contained in Z_{ij} and all analysts have the same information set. Setting $Z_i = Z_{ij}$ implies

$$F_j \pi_i = x_i + z_i,$$

which satisfies Assumption 1 by setting $\alpha = 1$ and $\eta_{ij} = 0$. Forecasts are unbiased with no noise.

Example 2: noisy information. Suppose for simplicity there is no public information so $x_i = 0$ and z_i is drawn from a Gaussian distribution with mean 0 and variance σ_z^2. Analysts receive noisy private signals of z_i, $s_{ij} = z_i + \nu_{ij}$, where ν_{ij} is a Gaussian noise with mean zero and variance σ_ν^2. Analysts use the distribution of z_i as their prior. Hence, an analyst's forecast is given by her posterior expectation:

$$F_j \pi_i = E(\pi_i | s_{ij}) = \underbrace{\frac{\sigma_z^2}{\sigma_z^2 + \sigma_\nu^2}}_{1-\lambda} s_{ij} = z_i \underbrace{-\lambda z_i}_{\equiv b_{ij}} + \underbrace{(1-\lambda)\nu_{ij}}_{\equiv \eta_{ij}}.$$

Setting $Z_i = \{z_i\}$, this model satisfies the first set of assumptions in Assumption 1, where $\alpha = (1-\lambda)$. Bias in this model is captured by λ: analysts underreact more when λ is bigger (i.e. when information is noisier). Noise comes from inference on noisy private signals $(1-\lambda)\nu_{ij}$.

Example 3: biased expectations, public information. Recently, a series of papers have suggested non-rational models of expectations formation (Bouchaud et al. 2019; Bordalo et al. 2019). In nearly all of these models, there is no soft information, so $Z_i = \varnothing$ and $z_i = 0$. For instance, Bordalo et al. (2019) suggest a "diagnostic" model with $X_i = \{X_{i0}, X_{i1}\}$. In this model, X_{i1} is diagnostic of π_i conditional on X_{i0}, but forecasters overreact to this information:

$$F \pi_i = E(\pi_i | X_{i1}) + \underbrace{\theta\left[E(\pi_i | X_{i1}) - E(\pi_i | X_{i0})\right]}_{\equiv b_i}.$$

This model satisfies Assumption 1 with no noise nor soft information. Bias is conditional on public information X_i. Similarly, a model with "sticky expectations" (as used in Bouchaud et al. 2019) where forecasters overweight the past realizations of an AR1 process, satisfies our assumptions. Like the "diagnostic" model, this model has only bias, but no noise.

2.2 Identifying the Decomposition

Given the assumptions in Assumption 1, we can now provide a version of the decomposition of MSE that we will be able to estimate.

Proposition 1. *Under Assumption 1, the generic decomposition of Lemma 3 writes:*

$$MSE^a - MSE^e = -\Theta + \left[\Delta + (1-\alpha)^2\Theta\right] + \frac{1}{J}\Sigma, \tag{8}$$

where

- $\Theta = E\left(z_i^2\right)$ *measures soft information;*
- $\Delta \equiv E\left[\left(E(\pi_i|X_i) - E(F\pi_i|X_i)\right)^2\right]$ *is the bias on public information;*
- $(1-\alpha)^2\Theta$ *is the bias on soft information;*
- $\Sigma \equiv var\left(\eta_{ij}\right)$ *is the individual expectation noise;*
- $\frac{1}{J} = E\left(\frac{1}{J_i}\right)$ *is the expected inverse number of forecasters per firm.*

Equation (8) writes our baseline decomposition (which always holds but cannot be estimated) as a function of the four key parameters that we will be able to estimate. Θ is the variance of z_i, i.e. the quantity of soft information a rational analyst would use. Total bias is the term in brackets, $\Delta + (1-\alpha)^2\Theta$. The first term is the bias on *public* information, and the second one is the bias on *soft* information: it increases with $(1-\alpha)^2$ and with the amount of soft information Θ. The final term is noise, $\frac{1}{J}\Sigma$. Since noise Σ is defined at the analyst level, it is inversely proportional to the number of forecasters because noise is independent across forecasters.

We now discuss how we identify Δ, α, Σ and Θ. First, note that Δ, the bias on public information, is directly identified from the data:

$$\Delta = E\left[\left(E(F_j\pi_i|X_i) - E(\pi_i|X_i)\right)^2\right].$$

This is the exercise most of the current literature on expectation bias is doing. The following proposition shows how the remaining three parameters are identified.

Proposition 2. *Define F_{ij}^* and π_i^* as residuals from projections onto observable information:*

$$F_{ij}^* \equiv F_j\pi_i - E\left(F_j\pi_i|X_i\right),$$

$$\pi_i^* \equiv \pi_i - E\left(\pi_i | X_i\right).$$

Under Assumption 1, α, Θ, and Σ are identified by the following moment conditions:

$$\begin{aligned} cov\left(\pi_i^*, F_{ij}^*\right) &= \alpha\Theta, \\ var\left(F_{ij}^*\right) &= \alpha^2\Theta + \Sigma, \\ cov\left(F_{ij}^*, F_{ik}^* \mid j \neq k\right) &= \alpha^2\Theta. \end{aligned}$$

The first moment condition in Proposition 2 states that the covariance between forecasts of analysts and realizations of EPS depends on soft information, Θ, and the weight analysts place on it, α. If analysts are unbiased on soft information, then this covariance directly estimates the size of soft information. The second condition simply states that analyst forecasts, after projecting out public information, can vary for two reasons: soft information and noise. Finally, the third moment condition is the covariance between forecasts of *different* analysts forecasting the *same* realization. Here our assumption that noise is uncorrelated across analysts is crucial, as it implies this covariance is due to analysts seeing the same soft information and sharing bias towards it.

To clarify identification, it is helpful to rewrite the three moment conditions in Proposition 2 as:

$$\begin{aligned} \alpha &= \frac{\text{cov}\left(F_{ij}^*, F_{ik}^* \mid j \neq k\right)}{\text{cov}\left(\pi_i^*, F_{ij}^*\right)}, \\ \Theta &= \frac{\text{cov}\left(\pi_i^*, F_{ij}^*\right)}{\alpha}, \\ \Sigma &= \text{var}\left(F_{ij}^*\right) - \text{cov}\left(F_{ij}^*, F_{ik}^* \mid j \neq k\right) \end{aligned}$$

The first equation shows bias on soft information is identified as the excess comovement of analyst forecasts, compared to the comovement of forecasts and realization. The intuition here is that if analysts are using their information correctly, their forecasts should have the same correlation as with realization. In contrast, if analysts excessively rely on soft information (e.g., $\alpha > 1$), their forecasts will be too correlated relative to comovement with the DGP. The second equation is straightforward: once we've identified α, Θ follows from rescaling the covariance between forecasts and realizations by analysts bias. Finally, Σ is identified as a residual variance: any variance in forecasts that cannot be accounted for by comovement between analysts. This is because noise across analysts is uncorrelated by assumption.

2.3 Estimation Strategy

We now discuss in detail how we use Proposition 2 in estimation. We start with public information, which is a separate block. The first step to estimate $E(\pi_i|X_i)$ and $E(F_j\pi_i|X_i)$ in order to compute Δ, the bias on public information, and to residualize forecasts and realization with respect to public information. To estimate these two conditional expectations, we pool all firms and periods (so that our simplified notation i corresponds to it in the application) and flexibly estimate $E(\pi_i|X_i)$ and $E(F_j\pi_i|X_i) = E(F\pi_i|X_i)$ using Elastic Net in one full-sample estimation. These estimations are performed separately for each forecasting horizon, h. We perform full-sample (rather than rolling) estimations because we are now focused on in-sample rather than out-of-sample fit and use Elastic Net because it is substantially less prone to over-fitting, given it has only one hyperparameter.[9] The subjective forecasts $F\pi_{it+h}$ we work with are consensus forecasts, consistent with Section 1.3 and the decomposition we are interested in (which is written at the consensus level).

Secondly, we use these ML estimates to residualize realizations and forecasts with respect to observables X_{it}. These residuals (F_{ij}^* and π_i^*) capture variation in earnings and analyst forecasts orthogonal to public information. We also use our ML estimates to calculate the public bias directly:[10]

$$\hat{\Delta} = \hat{E}\left[\left(\hat{E}(F\pi_i|X_i) - \hat{E}(\pi_i|X_i)\right)^2\right]$$

where $\hat{E}$ denotes the sample expectations estimated by our supervised learning estimators in the first step. By replacing conditional expectations with function approximations from our machine learning estimators, we are implicitly making the assumption that our machine learning estimators are consistent at reasonable rates given our sample size. Without placing further restrictions on the data-generating process for π_{it+h}, there are no theoretical results that justify this assumption. However, there is a growing theoretical literature suggesting this assumption is satisfied under a variety of reasonable assumptions about the DGP.[11]

Finally, estimate the three remaining parameters, Δ, Θ, and Σ by performing a separate GMM estimation with the three moment conditions in Proposition 2 for each h. Since we have separate estimates for each h, we index our estimates with h subscripts (e.g. Δ_h). Our estimation requires

[9] We obtain quantitatively similar results with our two tree-based methods, but choose present the results with Elastic Net for simplicity.

[10] Under the assumption that our machine learning estimators are consistent, this is formally justified by the continuous mapping theorem.

[11] For asymptotic results on the approximation error of various supervised learning estimators under different DGP assumptions in large samples, see Belloni et al. (2011) for Iterative Lasso, Chetverikov, Liao, and Chernozhukov (2020) for cross-validated Lasso, Wager and Athey (2018) for random forests, and Schmidt-Hieber (2020) for deep neural networks.

some reweighting to account for the fact most analysts forecasts non-overlapping sets of firms. We discuss the details of the GMM procedure in Appendix E.

2.4 Results

Parameter estimates. Figure 2 presents our estimates for the eight different forecasting horizons h. Red circles correspond to quarterly forecasts, while blue squares correspond to annual forecasts. Our estimates of $(\Delta_h, \Theta_h, \Sigma_h)$ are normalized by the realized mean of π^2_{it+h}. This is a natural normalization as these elements contribute to the MSE, which we normalize the same way (see Section 1.3).

Focusing first on quarterly forecast horizons (the red squares in Figure 2), we estimate a noise of around 5% and a public bias of around 4%. We can reject the null hypothesis of no deviations from rational expectations as these elements are statistically different from zero. However, at short-horizons these deviations are dominated by a large amount of soft information, consistent with the fact that analyst forecasts are overall more accurate than econometric forecasts at these horizons (Table 2). The amount of soft information decays however relatively quickly, as does the relative accuracy of analyst forecasts. Looking at α, we see that bias on soft information is non-zero but small. At shorter horizon, soft information is big but $\alpha \approx 1$. At longer horizons, $\alpha > 1$, but there is little soft information. For example, at the four-quarter horizon, $\alpha = 1.35$, which implies a soft bias of $0.35^2\Theta \approx 1\%$.

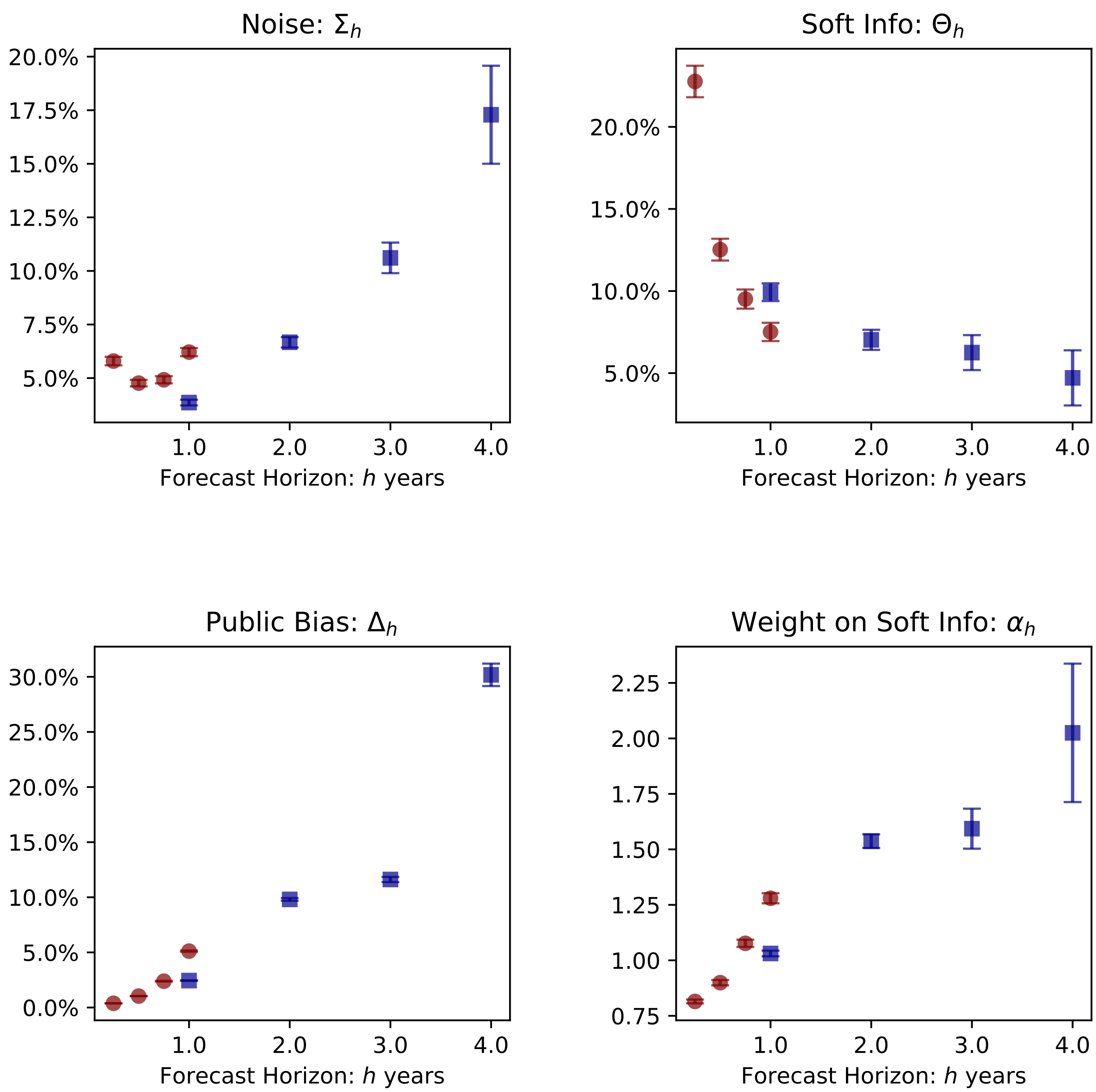
Turning to annual forecasts (the blue squares in Figure 2), we see the first main result of our paper. The term structure of noise and bias are upward-sloping. In contrast, the quantity of soft information decreases with horizon, first very fast, then at a slower pace after one year. For each additional year, we estimate the amount of noise in subjective forecasts increases by a factor of around 1.5-2. We find a similar pattern for public bias.[12]

A key contribution of our paper is to measure bias on soft information $(1 - \alpha_h)^2\Theta)h$, and to compare it with bias on public information Δ_h (which most of the literature focuses on). We provide this comparison in Figure 3. At short horizons, both biases are relatively small, and of comparable magnitude (a few percent of the mean sum of squared realizations). At longer horizons, observable bias becomes very large (30% of the mean sum of squared realizations), while the amount of bias on soft information remains modest. As we noted above, this is because there is not much soft

[12] One possible concern with these results is that the sample of firms changes across forecasting horizons. We have performed our estimation on a subsample of firm-years for which we have forecast data on all horizons. The slope of the term structures we estimate are quantitatively similar.

Figure 2. **The Term Structure of Information, Bias, and Noise**

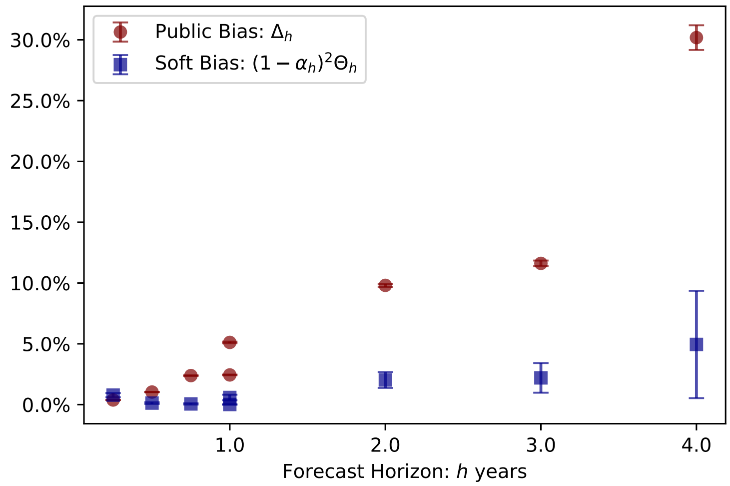
This figure plots our results from estimating our four parameters of interest across our eight forecasting horizons. Θ_h is the soft information; Δ_h is the analyst bias on public information; Σ_h is the analyst noise; $\alpha_h - 1$ is the bias on soft information. Red circles are quarterly forecasts and blue squares are annual forecasts. For the results plotted here, we used Elastic Net to estimate the two conditional expectation functions. All estimates (except for α_h) are normalized by the average squared earnings-per-share divided by price, calculated across this entire sub-sample for the relevant forecasting horizon. Error bars represent 95% confidence intervals. See Section 2.3 and Appendix E for additional details.



information at longer horizons.

Figure 3. **Bias on Public vs. Soft Information**

This figure plots our estimate of analyst bias on public and soft information in red circles and blue squares respectively. Both parameter estimates are normalized by the average squared earnings-per-share divided by price, calculated across this entire sub-sample for the relevant forecasting horizon. Error bars represent 95% confidence intervals. See Section 2.3 and Appendix E for additional details.



MSE decomposition. Overall, bias and noise increase at longer forecasting horizons, while soft information decays. These all could explain the fact from Section 1.3 that subjective forecasts lose their comparative advantage at longer horizons relative to statistical forecasts. Using the decomposition of Proposition 1, we can attribute the term structure of $MSE^a - MSE^e$ to each one of its components: $-\Theta$, Δ, $\Theta(1-\alpha)^2$, and $J^{-1}\Sigma$.

Table 3 reports the results, where (as before) each component is normalized by the mean sum of squared realizations. At short forecasting horizons such as one or two quarters, we find soft information is the main reason subjective forecasts outperform statistical forecasts: public bias, private bias, and noise combined are only around 10-20% of soft information. Within both sources of bias and noise, public bias plays the largest role at these short forecast horizons, but it is still small relative to soft information.

Consistent with Figure 2, the picture changes at longer forecasting horizons. Starting at four quarters, bias and noise combined are as large as soft information. At forecasting horizons of greater than one year, where econometric forecasts outperform subjective forecasts, bias and noise increase from around 1.5x soft information at two years to 3x at three years and over 9x at four years. Most of the ability of machine forecasts to outperform statistical forecasts is driven by public bias and, to a lesser extent, noise. Noise plays a smaller role because its effect is dampened by $\frac{1}{J}$. Private bias also plays a small role simply because the quantity of soft information declines at longer horizons.

Table 3. **Decomposition of** $MSE^a - MSE^e$

This table performs the decomposition of the difference between the mean squared error of consensus forecasts MSE^a, and that of econometric forecasts, MSE^e, shown in (8) at each forecasting horizon, h. All values are normalized by the average squared earnings-per-share divided by price, calculated across this entire sub-sample for the relevant forecasting horizon. MSE^a, MSE^a, and $\frac{1}{J}$ are calculated by taking sample expectations over a panel of firm-year-analyst pairs, which is their difference is slightly different from Table 2. See Appendix E for additional details.

Horizon: h	$MSE^a - MSE^e$	$-\Theta$	Δ	$(1-\alpha)^2\Theta$	$\frac{1}{J}\Sigma$
1 Quarters	-20.39%	-22.77%	0.38%	0.78%	1.21%
2 Quarters	-10.31%	-12.52%	1.03%	0.13%	1.05%
3 Quarters	-5.99%	-9.51%	2.39%	0.06%	1.08%
4 Quarters	-0.48%	-7.51%	5.12%	0.59%	1.32%
1 Years	-6.75%	-9.93%	2.44%	0.01%	0.73%
2 Years	6.15%	-7.03%	9.81%	2.03%	1.35%
3 Years	11.74%	-6.25%	11.61%	2.2%	4.18%
4 Years	41.23%	-4.71%	30.18%	4.95%	10.81%

In sum, our estimation in uncovers a strong upward-sloping term structure of bias and noise, but a decreasing – yet convex – term structure of soft information. In the remainder of the paper, we first discuss two implications of an upward-sloping term structure of noise in Section 3 and then examine what restrictions our estimates place on models of belief formation.

3 Implications of Expectation Noise

3.1 Interpreting the Coibion-Gorodnichenko Coefficient

The first implication of our decomposition is that it provides a simple explanation why the Coibion-Gorodnichenko (CG) coefficient should decrease and become negative at longer horizons. While this fact can be interpreted as increasing overreaction at longer horizon, we show more negative CG coefficient naturally emerges from the upward term structure of noise that we uncover.

The CG coefficient at horizon h is defined the slope coefficient, β_{CG}, in the following OLS regression:

$$\pi_{it+h} - F_t^j \pi_{it+h} = \alpha + \beta_{CG} \left(F_t^j \pi_{it+h} - F_{t-1}^j \pi_{it+h} \right) + e_{it}. \tag{9}$$

At an intuitive-level, β_{CG} measures over- and under-reaction. β_{CG} will vary with horizon h, but we omit the index to lighten notations. If $\beta_{CG} > 0$, updates predict positive (ex-post pessimistic) errors, so this is underreaction. When $\beta_{CG} < 0$, this is taken as evidence of overreaction.

Now, imagine forecasts are biased (say, over- or under-reacting), but also noisy. In this case, it is easy to see that noise will make the coefficient β_{CG} smaller (or even negative) that what the pure pattern of under- or over-reaction would suggest. Our estimation allows us to measure this bias in the data, and we find that it is very large, especially at long horizon where noise is big.

To see this, denote $\overline{F}_t^j \pi_{it+h} = F_t^j \pi_{it+h} - \eta_{ijt}^h$ as the forecast in a world without noise. Denote the variance of forecast revisions from the data is σ_{rev}^2 and $\overline{\sigma}_{\text{rev}}^2$ as the variance of noiseless forecasts. The following result characterizes the relationship between the observed CG coefficient (generated by noisy forecasts) and a counter-factual CG coefficient with pure bias and no noise.[13]

Proposition 3. *Assume the noise term at t, η_{ijt}^h, is uncorrelated with the noise term at $t-1$, η_{ijt-1}^h. Denote the CG coefficient estimated using observed forecasts at horizon h as β_{CG} and the CG coefficient estimated using noiseless forecasts as $\overline{\beta_{CG}}$. Then:*

$$\beta_{CG} = \overline{\beta_{CG}} * \frac{\overline{\sigma}_{rev}^2 - \Sigma_h}{\sigma_{rev}^2}$$
$$\sigma_{rev}^2 = \overline{\sigma}_{rev}^2 + \Sigma_h + \Sigma_{h+1}.$$

where Σ_h is the noise contained in forecasts of horizon h. Thus, given measures of noise, one can infer $\overline{\beta_{CG}}$ and $\overline{\sigma}_{rev}$ from β_{CG} and σ_{rev}.

Proposition 3 allows us to compute the noiseless $\overline{\beta_{CG}}$ and $\overline{\sigma}_{\text{rev}}$ from their observed counterparts. It also shows that noise has two effects on the CG coefficient. First, noise induces a negative correlation between forecast errors and forecast revisions, both of which contain the same noise term with opposite sign. The second effect is a classic attenuation bias: noise induces measurement error for the "true" revision, so the coefficient is smaller in absolute value.

In Figure 4, we show that noise obscures inference about over- or under-reaction based on the CG coefficient. First, we report the observed CG coefficient directly in raw data.[14] The results

[13]This result requires noise to be uncorrelated over time, as stated in the Proposition. We view this as a reasonable assumption since it holds in most formulations of noisy expectations models.

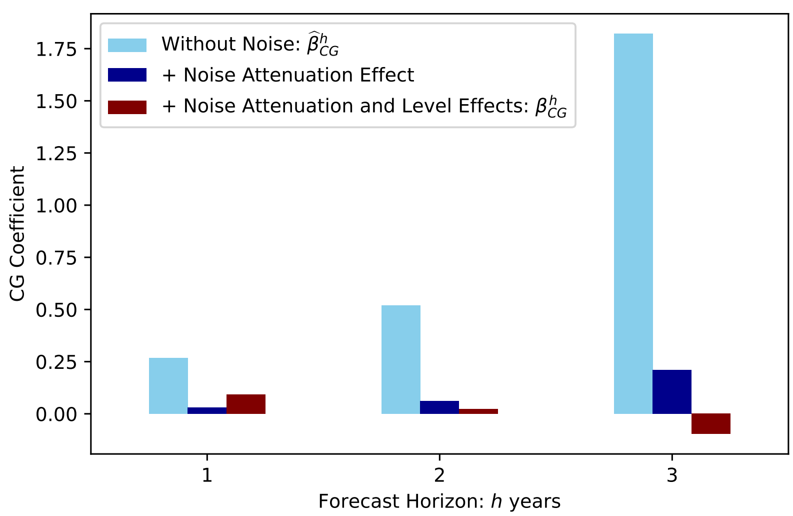
[14]We estimate the regression pooling all firm-year-analyst observations, as in Bouchaud et al. (2019).

of this regression are shown in the red bars in Figure 4. At the one-year horizon, we estimate $\beta_{CG} \approx 0.1$ (consistent with Bouchaud et al. 2019). At longer horizons, the CG coefficient decreases monotonically, flipping sign at the three-year horizon. This pattern is consistent with existing literature that finds individual-level forecasts tend to under-react at short-horizons (e.g., Bouchaud et al. 2019) and over-react at longer horizons (e.g., Giglio and Kelly 2018; Bordalo et al. 2020; D'Arienzo 2020).

Next, we use Proposition 3 to compute the noise-free $\overline{\beta_{CG}}$, which is shown in light blue bars in Figure 4. In contrast to observed CG coefficient, our estimate of the counter-factual noiseless CG coefficient is positive at all horizons and *increases* with the horizon. Thus, the bias contained in forecasts shows *much more underreaction* at longer horizons, instead of a flip to overreaction. The dark blue bar in the middle is the pure effect of attenuation, which cannot flip the sign of the CG coefficient, but is shown to reduce its magnitude considerably.

Figure 4. Noise and Coibion-Gorodnichenko Regression Coefficient

The figure shows the effect of expectation noise on the analyst-level CG coefficient. The x-axis in the figure is the forecasting horizon, h. For each h, we estimate the CG coefficient by estimating (9) on the sample of firm-years-analysts for which we have horizon h forecasts at t and horizon $h+1$ forecasts at $t-1$. This estimate is shown in the rightmost red bars, denoted by β_{CG}. We then compute the CG coefficient without noise using Proposition 3, $\overline{\beta}_{CG}$, which is shown in the leftmost light blue bars. In the middle blue bars, we plot the CG coefficient that would be obtained with just the noise attenuation effect, which is $\beta_{CG} + \sigma_{rev}^{-2}\Sigma$. To account for the fact that the sample for which we observe forecast revisions is smaller than our sample in Table 1, we reestimate the two noise terms required for each horizon on this subsample.



3.2 Forecast Complementarity

In addition to affecting the term structure of the CG coefficient, the results in Figure 2 highlight a tension from the perspective of a forecaster. On the one hand, subjective forecasts are biased and noisy at longer horizons. However, at the same time, subjective forecasts still contain non-trivial amounts of soft information, even at longer horizons. Perfect (unbiased) use of this soft information would result in about the same order of magnitude of gain in MSE from relying on analyst forecasts over econometric forecasts at short horizons (Table 2). Together, these findings suggest there should be some complementarity between subjective and statistical forecasts at longer horizons.

To examine the presence of such complementarity, we ask how much predictive power consensus forecasts add to the information set of the econometrician, X_{it}. Characterizing the MSE of this combined forecast is not possible without further distributional assumptions. To build intuition, the following proposition assumes that soft information and noise are both normally distributed.

Proposition 4. *Assume z_i and η_{ij} are jointly normally distributed. Then*

$$MSE^a - MSE^{e+a} = \left[\Delta + (1-\alpha)^2\Theta + (1-\beta^2)\frac{1}{J}\Sigma\right] - \left[(1-\alpha\beta)^2\Theta\right],$$

where $\beta = \frac{\alpha\Theta}{\alpha^2\Theta + \frac{1}{J}\Sigma} \leq 1$.

Proposition 4 explains how the combined forecasts compares to the pure analyst forecast. The first term in brackets captures the fact that econometric adjustment reduces noise and bias. It optimally gets rid of predictable bias. This is because one projects the forecast error on observables, and subtracts them from the subjective forecast. It also adjusts the bias on soft information $\Theta(1-\alpha)^2$, and reduces the amount of noise $\frac{1}{J}\Sigma$.

However, the second term in brackets term shows there is a trade-off: the combined forecast differs in terms of how it uses soft information, placing weight $\beta\alpha$ on soft information instead of α. This term arises because the combined forecast faces a trade-off when deciding how much weight to put on subjective forecasts (i.e. β): increasing the weight allows it to leverage soft information, but also introduces more noise. When there is no noise, $\beta = \alpha^{-1}$ and the combined forecast corresponds to FIRE.

In Table 4, we empirically explore the relative performance of the combined forecast and the analyst consensus. The layout is identical to Table 2: we show the MSEs at different forecast horizons along with Diebold-Marino test statistics under a squared loss function. At quarterly and the one-year horizons, we find the combined forecast cannot meaningfully beat the analyst consensus. This is to be expected, given our results in Figure 2. Subjective forecasts are not very biased at short horizons and have substantial soft information, which makes them a tough benchmark to beat ($\alpha \approx 1$). At longer horizons, however, the combined forecast dominates by a large amount: 9 and 21 percentage points of realized MSS at the three- and four-year horizons, respectively. However, comparing these results to those in Table 2 shows the improvement relative to pure econometric forecasts is small: around 1 to 3 percentage points. This is consistent with the trade-off highlighted Proposition 4: although long horizon subjective forecasts contain information, the upward-sloping term structure of noise means it's difficult to extract it.

Table 4. The Term Structure of Forecasting Accuracy: Analyst vs. Econometrician + Analyst

This table contains the mean squared error of analyst forecasts in the first column, denoted MSE_h^a, and of our econometrician + analyst forecasts, denoted MSE_h^{e+a}, across different forecasting horizons when forecasting the realization of EPS at $t+h$ divided by price-per-share at t. The numbers reported in the table are normalized by the mean squared realization of earnings-to-price at each horizon, which represents the percentage utility loss relative to having perfect foresight in the interpretive model presented in each year (Section 1.3). In parentheses, we report the Diebold-Marino test-statistics for testing the relative accuracy of the two forecasts under a squared loss function, where the asymptotic variance is calculated by performing a bootstrap at the year level with 1,000 iterations.

	MSE_h^a	MSE_h^{e+a}		
Horizon: h	Analyst	Elastic Net	Random Forest	Boosted Trees
1 Quarters	4.6%	4.56%	4.81%	4.71%
		(-2.28)	(6.64)	(3.97)
2 Quarters	8.43%	8.49%	8.9%	8.76%
		(1.3)	(6.6)	(5.57)
3 Quarters	13.05%	12.85%	13.02%	12.95%
		(-1.54)	(-0.19)	(-0.69)
4 Quarters	18.71%	17.61%	17.73%	17.69%
		(-4.05)	(-3.04)	(-3.35)
1 Years	9.9%	9.82%	10.01%	10.0%
		(-0.71)	(0.74)	(0.69)
2 Years	29.19%	25.49%	24.13%	25.13%
		(-5.32)	(-5.97)	(-4.5)
3 Years	33.32%	25.1%	24.37%	25.4%
		(-15.03)	(-13.69)	(-11.32)
4 Years	46.41%	26.73%	25.29%	27.55%
		(-17.36)	(-17.22)	(-16.39)

We conclude this section by noting an additional implication of Proposition 4 for optimal forecasting. Although combining subjective and statistical forecasts is no longer a "free lunch" in the presence of noise, increasing the number of forecasters in the consensus, J, is (as in Kahneman et al. 2021). This result follows from our assumption that noise is uncorrelated across forecasters, so noise in the consensus forecast will be averaged out as the number of forecasters increases. Changing the number of forecasters is clearly not possible in our setting, but this insight could be useful in organizational settings, such as forecasting demand (e.g., Bajari, Chernozhukov, Hortaçsu, and Suzuki 2020) or hiring workers (e.g., Bergman, Li, and Raymond 2020).

4 Models of the Term Structure of Bias and Noise

In this section, we explore the extent to which existing expectations models can *jointly* fit our estimated patterns in the term structure of bias and noise from Section 2. We first show that these models, in their simplest form, cannot jointly match the term structures of noise and biases. Thus, a mechanism needs to be added – we propose a simple model in the spirit of Patton and Timmermann (2010). This mechanism is portable and could be added to existing models. We then estimate this model and show it also can explain the cross-sectional relationship between noise and volatility.

4.1 Existing Models

We first consider a list of classic models for the term structure of bias and noise. We primarily focus on variants and extensions of noisy information models, which is a standard framework that has predictions on the term structure of expectation noise and bias.

Setup. Because horizon is a critical part of our discussion here, we revert to notation with explicit time of forecast t and horizon h. We omit i and j as these indices are not important in this discussion (so there is, say, only one analyst and one firm). Given our notations and key structural assumptions, the DGP and consensus forecast write:

$$\begin{aligned} \pi_{t+h} &= x_t^h + z_t^h + \varepsilon_t^h, \\ F_t \pi_{t+h} &= \underbrace{g_h(X_t)}_{=E(F_t \pi_{t+h}|X_t)} + \alpha_h z_t^h + \eta_t^h. \end{aligned} \tag{10}$$

Each model we will consider delivers a forecasting equation of the form in (10).

In line with the literature, we also impose additional structure on the data-generating process. We do this in this Section only, and mostly in order to clarify the discussion. The structure we impose on the DGP is described below:

Assumption 2. *The laws of motion for x_t and z_t are*

$$x_t = \rho_x x_{t-1} + u_t^x, \quad z_t = \rho_z z_{t-1} + u_t^z,$$

where $E\left(u_t^x \mid x_{t-1}\right) = E\left(u_t^z \mid z_{t-1}\right) = 0$ *and* $[\rho_x, \rho_z] \in (0,1)^2$.

4.1.1 Baseline Noisy Information Model

We first consider a baseline noisy information model in the spirit of Woodford (2003). It is the most natural starting place because it generates both bias (from the viewpoint of the econometrician) and noise, as shown in Section 2.1. In this model, the econometrician observes x_t but not z_t. The analyst observes noisy signals of x_t and z_t, denoted by $\mathcal{S}_t^x$ and $\mathcal{S}_t^z$, respectively. Given these signals, the analyst applies Bayes rule to form forecasts as follows:

$$F_t x_t = E\left(x_t \mid \mathcal{S}_t^x\right), \quad F_t z_t = E\left(z_t \mid \mathcal{S}_t^z\right), \tag{11}$$

$$F_t x_{t+h} = \rho_x^{h-1} F_t x_t, \quad F_t z_{t+h} = \rho_z^{h-1} F_t z_t, \quad F_t \pi_{t+h} = F_t x_{t+h} + F_t z_{t+h}. \tag{12}$$

Equations (11) and (12) characterize the two-step process performed by the forecaster in the noisy information model. First the forecaster forms beliefs about the current state, which is rational conditional on her information set. Next, the forecaster forms h-period-ahead forecasts by combining her knowledge of the data generating process with her forecasts from the first step.

This model is enough to pin down the term structure of bias and noise. The following proposition summarizes the results.

Proposition 5. *Denote Θ_h, Σ_h, Δ_h and α_h soft information, noise, public bias, and soft bias at horizon h. Then, in the baseline noisy information model, the term structure of public and soft bias are downward sloping:*

$$\frac{\Delta_{h+1}}{\Delta_1} = \rho_x^{2h} \leq 1$$

$$\frac{(1-\alpha_{h+1})^2 \Theta_{h+1}}{(1-\alpha_1)^2 \Theta_1} = \rho_z^{2h} \leq 1.$$

The term structure of noise is also downward sloping:

$$\frac{\Sigma_{h+1}}{\Sigma_1} = \theta \rho_x^{2h} + (1-\theta)\rho_z^{2h} \leq 1,$$

where $\theta \in [0,1]$ is the fraction of total noise at $h = 1$ that comes from $\mathcal{S}_t^x$.

To build intuition for this result, assume $h \to \infty$. Because x_t and z_t are stationary process, the best infinite horizon forecast is their long-run mean, 0. Thus, the analyst will be unbiased in this extreme case and will also issue noiseless forecasts because she will place no weight on her sequence of noisy signals. Thus, bias and noise should decline at long horizon. The evidence in Section 2.4 provides a clear rejection of this prediction, as the term structures of noise and bias are upward sloping.

4.1.2 Variants of Baseline Noisy Information Model

We now discuss the predictions of commonly-used variants of the noisy information model for the term structure of public bias and noise.

Bounded rationality. A common microfoundation for noisy information models is bounded rationality (e.g., Sims 2003). In these models, the set of signals is endogenously chosen to maximize an objective function decreasing in forecast errors, subject to a cost function increasing in the mutual information of the signals. Although these models introduce a tight connection between the signals and primitives (e.g., signal precision and cognitive capacity), they also have downward sloping term structures of bias and noise because they satisfy equations (11) and (12).

Diagnostic expectations. Bordalo et al. (2020) combine diagnostic expectations, which generates overreaction to recent news, with noisy information. This model breaks (11) due to nonrational expectations about the current state. But the dynamics of forecast are an AR1 as in equation (12). Thus, this model exhibits a downward sloping term structure of bias and noise for the same reason as the baseline noisy information model. Intuitively, the analyst knows x_t is mean-reverting, so even if she overreacts to news at short horizons, she knows that in the long-run, x_t will go back to the long-run mean.

Over-confidence. Another common way to generate non-rational reactions in the noise information framework is via agents' overconfidence about their signal qualities (e.g., Daniel et al. 1998; Eyster, Rabin, and Vayanos 2019). In the common case where $\mathcal{S}_t^x$ consists of signals each period with independent normal errors, over-confidence corresponds to the analyst updating using a variance lower than the true signal variance. Since overconfidence only changes (11), it will only change the level of bias and noise, but not the term structure of forecasts.

4.1.3 Alternative Frameworks that Break (12)

The previous section shows noisy information models cannot match the term structure of bias and noise because (12) holds: forecasters are rational in the long-run, because they know the DGP's parameters. We now consider three sets of models that break this equation in different ways.

Learning about the mean. The second way to break (12) is to assume the forecaster believes the long-run mean of x_t is $\hat{\mu}_x \neq \mu_x$. Afrouzi et al. (2021) provide a model of this sort, where the forecaster does not know μ and consequently overweights recent information in her estimation. In Appendix F, we show this model has the potential to qualitatively match our data, as long-run forecasts are further from the rational expectation, but cannot do so quantitatively.

Misspecified stationary model. The first way to break (12) is to assume the forecaster believes the persistence of x_t is $\hat{\rho}_x > \rho_x$, while maintaining noisy information. Angeletos et al. (2020) show this over-extrapolation is necessary for matching over-reaction in macroeconomic expectations. Although this type of over-extrapolation can make the term structure of bias and noise less downward sloping in our setting, it remains downward sloping, because the forecaster still believes that the process mean-reverts.

Misspecified non-stationary model. Fuster, Laibson, and Mendel (2010) propose a framework in which the law of iterated expectations fails, known as natural expectations. In this framework, the true DGP is a stationary AR(2) in levels, but the forecaster has an "intuitive" DGP that is an AR1 in changes. Importantly, because the intuitive model is non-stationary, this model gets a larger weight at longer horizons, which generates an upward-sloping term structure of bias. But this model does not have noise, which is why we turn to the following structure.

4.2 Proposed Model

Model description. Consistent with our approach in Section 2, we deviate from prior literature and place no restrictions on the data generating process. We assume that forecasts are described by:

$$\begin{aligned} F_t \pi_{t+h} &= (1-m_h) d_t^h + m_h E(\pi_{t+h} \mid X_t, Z_t), \\ &= (1-m_h) d_t^h + m_h \left(x_t^h + z_t^h \right). \end{aligned} \tag{13}$$

This forecasting equation is motivated by the inattention framework of Gabaix (2014), where d_t^h corresponds to a cognitive default and $m_h \in [0,1]$ represents the horizon-specific amount of attention to processing the available information. If $m_h = 1$, the analyst issues a forecast equal to the conditional expectation. Unlike Gabaix (2014) however, and in the spirit of Patton and Timmermann (2010), we let d_t^h be a random variable, which we parameterize as follows:

$$d_t^h = \beta_0 + \beta_x x_t^h + \beta_z z_t^h + \nu_t^h, \quad \text{var}(\nu_t^h) \equiv \sigma_\nu^2.$$

The cognitive default, d_t^h, may contain noise, ν_t, which we assume is independent across horizons, forecasters, and firms. We also allow for the default to potentially depend on the state variables via β_x and β_z. The fact that d_t^h is a *noisy default* is crucial to this framework. Note that this model nests the case of full-information rational expectations if $m_h = 1$, and the case of unbiased but noisy forecasts if $\beta_x = \beta_z = 1$ and $\beta_0 = 0$.

Like in Patton and Timmermann (2010), we discipline the term structure of m_h through one parameter only, and assume

$$m_h(\kappa) = \frac{\kappa^2}{\kappa^2 + MSE_h^{FIRE}}, \tag{14}$$

where

$$MSE_h^{FIRE} = MSE_h^e - \Theta_h = E\left[\left(\pi_{t+h} - E\left(\pi_{t+h} \mid X_t, Z_t\right)\right)^2\right]$$

is the MSE of the full information rational expectations (FIRE) forecast. This specification of m_h captures of a form of bounded rationality: the forecaster relies more on her default when π_{t+h} is harder to predict. It will be key to matching the upward-sloping term structure of noise and bias. At long horizons, profits are harder to predict, so forecasters will lean more on noisy and biased defaults.

Term structures of noise and bias In this model, we can derive the term structure of bias and noise, summarized in the following proposition.

Proposition 6. *Public bias, soft bias, and noise are given by:*

$$\begin{aligned}
\Delta_h &= (1-m_h)^2 E\left[\left(\beta_0 + (\beta_x - 1)x_t^h\right)^2\right], \\
\alpha_h &= 1 + (1-m_h)(\beta_z - 1), \\
\Sigma_h &= (1-m_h)^2 \sigma_\nu^2,
\end{aligned}$$

The level of public bias is determined by the extent which the default is unconditionally biased (β_0), and biased on x_t^h, captured by β_x being different from 1. Also, this bias is magnified by the reliance on default (large $(1-m_h)^2$). Similarly, the bias on soft information is larger if the forecaster relies on the default more (m_h small) or if the default is more biased ($\beta_z - 1$ larger). Lastly, expectation noise comes from default noise, σ_v, and the extent to which the forecaster relies on the default, m_h.

The expressions in Proposition 6 illustrate how our model links bias and noise. In order to match the upward-sloping term structures of noise and bias, we need m_h to decrease in h. This will not be hard to achieve, since MSE_h^{FIRE} increases with h. The question is whether the model will be quantitatively succesful.

In sum, our model has only five parameters: (i) σ_v^2, which pins down by the average level of noise in subjective forecasts; (ii) β_0, β_x, and β_z, which pin down the amount of public and soft bias; (iii) κ, which determines the relative weight of the rational forecast and governs the term structure of both noise and bias.

Estimation. We estimate the model using a minimum distance estimator in which we match model moments to their empirical counterparts. In addition to the term structures of bias (public and soft) and noise, we target two other moments: the intercept and slope coefficient from regressing $F_t\pi_{t+h}$ onto x_t^h, which we denote as δ_0 and δ_1 respectively. These moments help identify β_0 and β_x. Denote $\theta = (\sigma_v, \beta_0, \beta_x, \beta_z, \kappa)$ the vector of the five model parameters. We define the vector of differences between model and empirical moments, $M_h(\theta)$, as:

$$M_h(\theta) = \begin{pmatrix} \alpha_h - [1-m_h(\kappa)]\beta_z - m_h(\kappa) \\ \Delta_h - [1-m_h(\kappa)]^2 E\left[\left(\beta_0 + (\beta_x - 1)x_t^h\right)^2\right] \\ \Sigma_h - [1-m_h(\kappa)]^2 \sigma_v^2 \\ \delta_h^0 - [1-m_h(\kappa)]\beta_0 \\ \delta_h^1 - [1-m_h(\kappa)]\beta_x - m_h(\kappa) \end{pmatrix}.$$

Given that we have four annual forecast horizons, these five term structures give us a total of $5 \times 4 = 20$ moment conditions that we stack into a single vector $M(\theta)$. We can then estimate the model using a minimum distance estimator. Given the scale difference of these moments, we weight this estimator with the inverse of the diagonal of the variance matrix from our GMM estimation in Section 2.4.

4.3 Estimation Results and Model Fit

The results from our minimum distance estimation are presented in Table 5. We find evidence of significant noise in the default: $\sigma_\nu \approx 0.06$. For reference, $\text{sd}(x_t + z_t) \approx 0.07$, implying the noisy we estimate in the default is around 0.85 times the noise in the data generating process. This significant cognitive noise is needed to match the average level of noise across the four forecasting horizons.

Our model's most important parameter is κ, which determines the *slope* of the term structure of bias and noise through m_h. We estimate $\kappa \approx 0.06$. To interpret this estimate, Table 5 presents the implied values of m_h from this value of κ, which range from $m_1 = 0.87$ to $m_4 = 0.53$.

Finally, our model has three parameters that control the cognitive default. We estimate $\beta_0 \approx 0.06$, $\beta_x \approx 0.77$, and $\beta_z \approx 2.1$. These estimates imply that there is fixed optimism in analyst forecasts and that the upward-sloping term structure of α_h is driven by analysts' over-weighting soft information and under-weighting public information, relative to the rational expectation.

Table 5. Minimum Distance Estimation Results

This table presents the results from estimating the five parameters of the model in Section 4.2 using minimum distance estimation. We target the term structures of α_h, Δ_h and Σ_h from Figure 2, in addition to the intercept and slope coefficients from regressing $F_t^j \pi_{it+h}$ onto x_{it}^h. This results in a total of 20 moments across our four annual forecast horizons. As a weighting matrix, we use the inverse of the diagonal of the variance matrix from our GMM estimation in Section 2.4. Standard errors are calculated using the covariance matrix from the GMM estimated in Section 2.4 and two-sided finite differentiation to calculate the Jacobian of the moment conditions. The table also shows the implied values of m_h, which are calculated by plugging the estimated value of κ into (14).

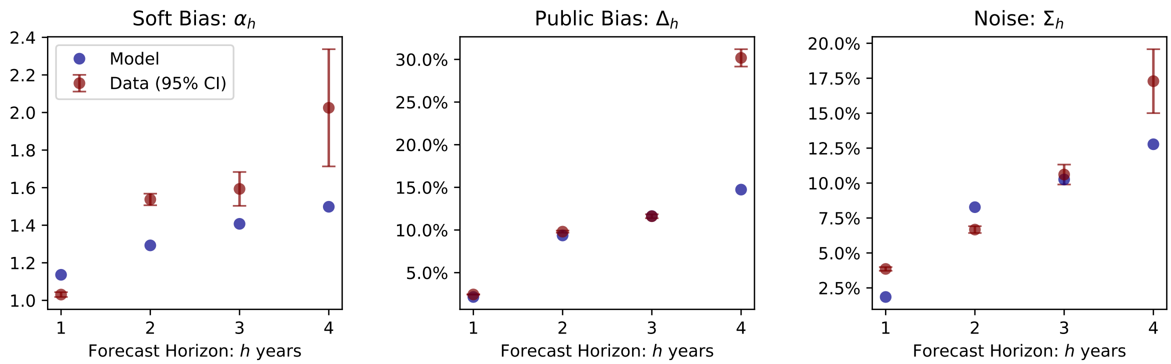
	κ	σ_ν	β_0	β_x	β_z
MDE Estimate	0.0561	0.0573	0.0648	0.7700	2.0730
Std. Error	(0.0001)	(0.0009)	(0.0009)	(0.0071)	(0.0400)
Implied m_h					
	$m_1 = 0.873$	$m_2 = 0.727$	$m_3 = 0.620$	$m_4 = 0.536$	

Figure 5 plots the term structures in our estimated model versus the data. The results show our model reproduces all three term structures reasonably well due to the decreasing term structure of m_h in Table 5. However, it struggles to fit the amount of bias and noise at long horizons, which is not entirely surprising because the function $m_h(\kappa)$ is concave in the rational forecast error at long horizons. Nevertheless, we view the fact that a model with one parameter, κ, controlling the slope of all three structures fits the data quite well as an important takeaway from this estimation.

These results suggest the underlying mechanisms generating bias and noise are linked, echoing the findings of Enke and Graeber (2020).

Figure 5. The Term Structure of Information, Bias, and Noise: Model and Data

This figure presents a comparison of the term structures of bias and noise in the data versus in the model in Section 4.2. The moments in the model are calculated using Proposition 6 and the parameter estimates in Table 5. The sample used here is the same sample as in Figure 2.



4.4 Noise and Volatility

A central ingredient in our model is that forecasters dealing with volatile processes tend to lean on their default more, which implies more bias and more noise: equation (14). In this section, we provide two additional pieces of evidence in support for this central ingredient: (i) noise is higher when volatility of the underlying process is higher and (ii) our model does a reasonable job reproducing this relationship out-of-sample.

We begin by examining how noise varies cross-sectionally with the volatility of the underlying DGP, for which we proxy using equity volatility. Ideally, we would measure volatility of earnings directly, but this is not possible because we have one annual earnings observation for each firm-year. We instead use daily equity return volatility because it is (1) easy to compute at the yearly level, and (2) a natural proxy for the volatility of cash flows, given most of the variation in firm-level valuation ratios is driven by cash flow news (Vuolteenaho 2002).

To perform this analysis, we split the sample into ten equally-spaced bins of equity volatility (computed using trailing 5-year windows of monthly returns). In each one of these bins, we then estimate (Σ, Θ, α), by applying the estimation strategy described in Section 2.3.[15] We thus come up

[15] The patterns in Figure 6 are robust to using different windows for calculating returns.

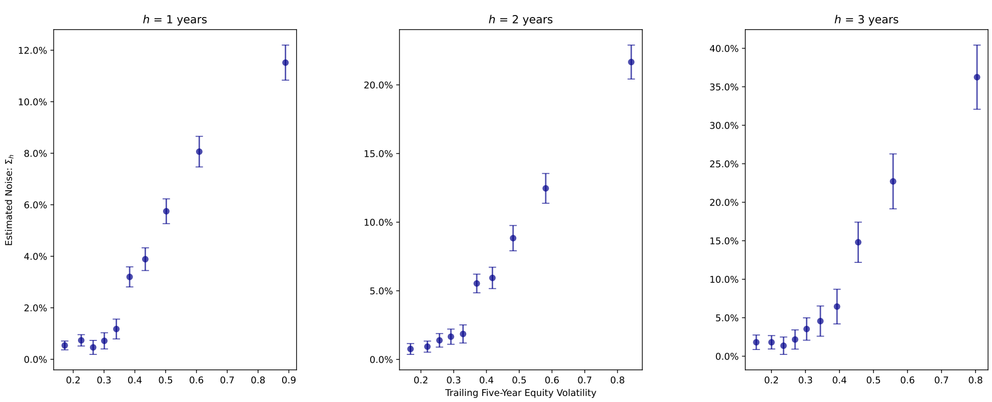
with 10 different values of noise Σ, which we normalize as usual by the sum of squared realizations. We then reproduce this procedure separately for three horizons: $h = 1, 2, 3$ years.

Panel A of Figure 6 shows that estimated noise increases almost monotonically in volatility. This relationship holds for all horizons. What is notable here is that our procedure to estimate Σ does *not* target returns or profit volatility. A conjecture is that high volatility bins correspond to high variance of forecast residuals, which then leads to larger noise.

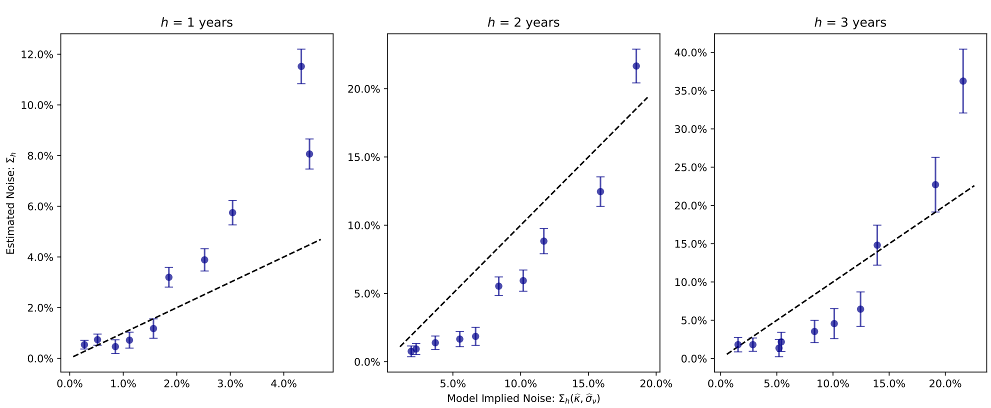
Figure 6. Noise and Volatility

Panel A of this figure plots our estimates of noise at three different horizons across evenly-spaced bins of equity volatility, measured using the annualized standard deviation of monthly stock returns from CRSP over the past five years for firm i in year t. Panel B of this figure plots this same empirical (conditional) noise against the model-implied noise for each bin. To estimate the model-implied noise in Panel B, we follow the procedure described in Section 4.4 to generate an estimate of MSE_h^{FIRE} within each bin, which we convert into an estimate of model implied noise using (14), Proposition 6, and our estimates of κ and σ_v from Table 5. Estimates are normalized by the mean squared EPS across the entire sample for each horizon h. Error bars represent 95% confidence intervals from the GMM estimation.

Panel A: Noise versus Equity Volatility



Panel B: Empirical versus Model-Implied Noise



We then check whether the model estimated the previous section is able to quantitatively predict the cross-sectional variation in noise. Our model predicts that noise is given by:

$$\Sigma_h = \left(\frac{MSE_h^e - \Theta_h}{\kappa^2 + MSE_h^e - \Theta_h} \right)^2 \sigma_v^2, \tag{15}$$

where we have used $MSE_h^{FIRE} = MSE_h^e - \Theta_h$. Thus, for each bin of volatility, we can use this formula to predict the amount of noise. We use the κ and σ_v estimated in the previous section, Θ_h estimated in the procedure above, and MSE_h^e from the data.

Panel B of Figure 6 compares this model-predicted noise with the observed noise (the one shown in Panel A). For each horizon, we show the value of noise predicted by our model on the x-axis and the empirically-estimated value of noise of the y-axis. If our model was correct, we would expect all points to lie on the 45-degree line (shown in black). The results show our model is not quantitatively off, but is statistically rejected at all three-horizons. Nevertheless, we view this quantitative fit as non-mechanical, given our estimation in Table 5 only targeted aggregate moments at different forecasting horizons to obtain κ and σ_v, while noise Σ is measured directly for each bin.

5 Conclusion

We find that subjective forecasts perform better than statistical forecasts at short horizons, but underperform at longer horizons. This decreasing relative accuracy of subjective forecasts is driven by an upward-sloping term structure of bias and noise, while the information advantage of subjective forecasters declines with horizon. Quantitatively the amount of noise we estimate at longer horizons is large: it generates a reversal in a commonly-used measure of over- and under-reaction.

Existing models, in their current form, lack a feature to match these upward-sloping term structures of bias and noise. We propose such a mechanism based on bounded rationality and noisy defaults. This model quantitatively matches these term structures. The model is parsimonious, as it has three key parameters: default noise, expectations bias, and the relative weight between rational forecast and noise default. This last parameter succeeds at matching both the term structures of noise and bias, suggesting a connection between the two. This model predicts that noise should be an increasing function of earnings volatility, a feature borne out by the data, both qualitatively and quantitatively.

Our model provides the reduced-form representation that a more micro-founded model should admit in order to match our empirical results. Subsequent work could enrich the model in this direction.

References

Afrouzi, Hassan, Spencer Kwon, Augustin Landier, Yueran Ma, and David Thesmar (2021), "New Experimental Evidence on Expectation Formation." *Working Paper*, 1–67.

Angeletos, George-Marios, Zhen Huo, and Karthik A. Sastry (2020), "Imperfect Macroeconomic Expectations: Evidence and Theory." *NBER Macroeconomics Annual*.

Bajari, Patrick, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki (2020), "The Impact of Big Data on Firm Performance: An Empirical Investigation." *Working Paper*.

Ball, Ryan T. and Eric Ghysels (2018), "Automated Earnings Forecasts: Beat Analysts or Combine and Conquer?" *Management Science*, 64, 4936–4952.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2011), "Inference for high-dimensional sparse econometric models." *Advances in Economics and Econometrics: Tenth World Congress Volume 3, Econometrics*, 245–295.

Bergman, Peter, Danielle Li, and Lindsey Raymond (2020), "Hiring as Exploration." *Working Paper*.

Bianchi, Francesco, Sydney C. Ludvigson, and Sai Ma (2020), "Belief Distortions and Macroeconomic Fluctuations." *Working Paper*.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer (2016), "Stereotypes." *The Quarterly Journal of Economics*, 1753–1794.

Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer (2019), "Diagnostic Expectations and Stock Returns." *Journal of Finance*, 74, 2839–2874.

Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer (2022), "Expectations of Fundamentals and Stock Market Puzzles." *Working Paper*.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer (2020), "Overreaction in Macroeconomic Expectations." *American Economic Review*, 110, 2748–2782.

Bouchaud, Jean Philippe, Philipp Krüger, Augustin Landier, and David Thesmar (2019), "Sticky Expectations and the Profitability Anomaly." *Journal of Finance*, 74, 639–674.

Bradshaw, Mark T., Michael S. Drake, James N. Myers, and Linda A. Myers (2012), "A re-examination of analysts' superiority over time-series forecasts of annual earnings." *Review of Accounting Studies*, 69–76.

Brown, Lawrence D., Andrew C. Call, Michael B. Clement, and Nathan Y. Sharp (2015), "Inside the "Black Box" of sell-side financial analysts." *Journal of Accounting Research*, 53, 1–47.

Brown, Lawrence D. and Michael S Rozeff (1978), "The Superiority of Analyst Forecasts as Measures of Expectations: Evidence from Earnings." *Journal of Finance*, 33, 1–16.

Bryzgalova, Svetlana, Jiantao Huang, and Christian Julliard (2020), "Bayesian Solutions for the Factor Zoo: We Just Ran Two Quadrillion Models." *Working Paper*.

Cao, Kai and Haifeng You (2020), "Fundamental Analysis via Machine Learning." *Working Paper*.

Chen, Qi and Wei Jiang (2006), "Analysts' weighting of private and public information." *Review of Financial Studies*, 19, 319–355.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney K. Newey, and James M. Robins (2016), "Double/Debiased Machine Learning for Treatment and Causal Parameters." *Working Paper*.

Chetverikov, Denis, Zhipeng Liao, and Victor Chernozhukov (2020), "On cross-validated Lasso in high dimensions." *Annals of Statistics*, 40.

Coibion, Olivier and Yuriy Gorodnichenko (2015), "Information rigidity and the expectations formation process: A simple framework and new facts." *American Economic Review*, 105, 2644–2678.

Daniel, Kent, Avanidhar Subrahmanyam, and David A. Hirshleifer (1998), "Investor Psychology and Security Market Under and Overreactions." *Journal of Finance*, 53, 1839–1885.

D'Arienzo, Daniele (2020), "Maturity Increasing Over-reaction and Bond Market Puzzles." *Working Paper*.

De la O, Ricardo and Sean Myers (2021), "Subjective Cash Flow and Discount Rate Expectations." *Journal of Finance*, 76, 1339–1387.

Dessaint, Olivier, Thierry Foucault, and Laurent Frésard (2020), "Does Big Data Improve Financial Forecasting? The Horizon Effect." *Working Paper*.

Enke, Benjamin and Thomas Graeber (2020), "Cognitive Uncertainty." *Working Paper*.

Eyster, Erik, Matthew Rabin, and Dimitri Vayanos (2019), "Financial Markets Where Traders Neglect the Informational Content of Prices." *Journal of Finance*, 74, 371–399.

Fuster, Andreas, David Laibson, and Brock Mendel (2010), "Natural Expectations and Macroeconomic Fluctuations." *Journal of Economic Perspectives*, 24, 67–84.

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

Gabaix, Xavier and David Laibson (2017), "Myopia and Discounting." *Working Paper*, 1–43.

Gershman, Samuel J. and Rahul Bhui (2020), "Rationally inattentive intertemporal choice." *Nature Communications*, 11.

Giglio, Stefano and Bryan T. Kelly (2018), "Excess volatility: Beyond discount rates." *The Quarterly Journal of Economics*, 133, 71–127.

Greenwood, Robin and Andrei Shleifer (2014), "Expectations of returns and expected returns." *Review of Financial Studies*, 27, 714–746.

Gu, Shihao, Bryan T. Kelly, and Dacheng Xiu (2018), "Empirical Asset Pricing via Machine Learning."

Hansen, Jorge W and Christoffer Thimsen (2020), "Forecasting Corporate Earnings with Machine Learning." *Working Paper*.

Harford, Jarrad, Feng Jiang, Rong Wang, and Fei Xie (2019), "Analyst career concerns, effort allocation, and firms' information environment." *Review of Financial Studies*, 32, 2179–2224.

Horowitz, Joel L. (2001), "The Bootstrap." *Handbook of Econometrics*, 5, 3159–3228.

Juodis, Artūras and Simas Kucinskas (2019), "Quantifying Noise." *SSRN Electronic Journal*, 2019, 1–61.

Kacperczyk, Marcin, Stijn Van Nieuwerburgh, and Laura Veldkamp (2016), "A Rational Theory of Mutual Funds' Attention Allocation." *Econometrica*, 84, 571–626.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein (2021), *Noise*. Little, Brown Spark, New York, Boston, and London.

Khaw, Mel Win, Ziang Li, and Michael Woodford (2019), "Cognitive Imprecision and Small-Stakes Risk Aversion." *Working Paper*.

Kothari, S. P., Eric C. So, and Rodrigo Verdi (2016), "Analysts' Forecasts and Asset Pricing: A Survey." *Annual Review of Financial Economics*, 1–23.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh (2020), "Shrinking the cross-section." *Journal of Financial Economics*, 135, 271–292.

Kumar, Alok, Ville Rantala, and Ruoxi Xu (2021), "Social Learning and Analyst Behavior." *Journal of Financial Economics*.

Maćkowiak, Bartosz and Mirko Wiederholt (2009), "Optimal sticky prices under rational inattention." *American Economic Review*, 993, 769–803.

Mankiw, N. Gregory and Ricardo Reis (2002), "Sticky information versus sticky prices: A proposal to replace the new Keynesian Phillips curve." *Quarterly Journal of Economics*, 117, 1295–1328.

Manski, Charles F. (2017), "Survey Measurement of Probabilistic Macroeconomic Expectations: Progress and Promise." *NBER Macroeconomics Annual*.

Mullainathan, Sendhil and Jann Spiess (2017), "Machine learning: An applied econometric approach." *Journal of Economic Perspectives*, 31, 87–106.

Nagel, Stefan (2021), *Machine Learning in Asset Pricing*. Princeton University Press, Princeton and Oxford.

Patton, Andrew J. and Allan Timmermann (2010), "Why do forecasters disagree? Lessons from the term structure of cross-sectional dispersion." *Journal of Monetary Economics*, 57, 803–820.

Patton, Andrew J. and Michela Verardo (2012), "Does beta move with news? Firm-specific information flows and learning about profitability." *Review of Financial Studies*, 25, 2789–2839.

Satopää, Ville, Marat Salikhov, Philip E. Tetlock, and Barb Mellers (2020), "Bias, Information, Noise: The BIN Model of Forecasting." *Working Paper*.

Schmidt-Hieber, Johannes (2020), "Nonparametric regression using deep neural networks with relu activation function." *Annals of Statistics*, 48, 1875–1897.

Sims, Christopher A. (2003), "Implications of rational inattention." *Journal of Monetary Economics*, 50, 665–690.

So, Eric C. (2013), "A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts?" *Journal of Financial Economics*, 108, 615–640.

van Binsbergen, Jules H., Xiao Han, and Alejandro Lopez-Lira (2020), "Man vs. Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases." *Working Paper*.

Vuolteenaho, Tuomo (2002), "What drives firm-level stock returns?" *Journal of Finance*, 57, 233–264.

Wager, Stefan and Susan C. Athey (2018), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association*, 113, 1228–1242.

Woodford, Michael (2003), "Imperfect Common Knowledge and Monetary Policy." In *Knowledge, Information, and Expectations in Modern Macroeconomics: In Honor of Edmund S. Phelps*, 25–58, Princeton University Press, Princeton, NJ.

Woodford, Michael (2020), "Modeling imprecision in perception, valuation, and choice." *Annual Review of Economics*, 12, 579–601.

INTERNET APPENDIX

Appendix A. Additional Derivations and Proofs

In this appendix, we provide derivations of the results stated in the main text. Recall that indices t and h have been suppressed, but all variables implicitly vary across years t and forecasting horizon h.

Proof of Lemma 1. Defining $\varepsilon_i = \pi_i - E(\pi_i \mid X_i, Z_i)$, the first equation holds trivially. The conditional mean independence conditions, $E(\varepsilon_i \mid x_i, z_i) = 0$ and $E(z_i \mid x_i) = 0$, follow from the law of iterated expectations. Starting with the second equality, we have

$$\begin{aligned} E(z_i \mid x_i) &= E\left[E\left(z_i \mid x_i, X_i\right) \mid x_i\right] \\ &= E\left(E(z_i \mid X_i) \mid x_i\right). \end{aligned}$$

From the definition of z_i, we have $E(z_i \mid X_i) = 0$. Combined with the previous expression, this delivers the desired result. The proof for the second equality follows similarly from the law of iterated expectations.

$$\begin{aligned} E(\varepsilon_i \mid x_i, z_i) &= E\left[E\left(\varepsilon_i \mid x_i, X_i, z_i, Z_i\right) \mid x_i, z_i\right] \\ &= E\left[E\left(\varepsilon_i \mid X_i, Z_i\right) \mid x_i, z_i\right] \end{aligned}$$

where the inner expectation equals zero by the definition of ε_i, which delivers the desired result. $\square$

Proof of Lemma 2. Defining $\eta_i = \pi_i - E(F^j \pi_i \mid X_i, Z_i)$, (3) follows trivally. The conditional mean independence condition, $E(\eta_{ij} \mid x_i, z_i) = 0$ follows from the law of iteration expectations.

$$\begin{aligned} E(\eta_{ij} \mid x_i, z_i) &= E\left[E\left(\eta_{ij} \mid x_i, X_i, z_i, Z_i\right) \mid x_i, z_i\right] \\ &= E\left[E\left(\eta_{ij} \mid X_i, Z_i\right) \mid x_i, z_i\right] \end{aligned}$$

where the inner expectation equals zero by the definition of η_{ij}, which delivers the desired result. $\square$

Proof of Lemma 3. Using Lemma 1, we have $MSE^e = E\left[(x_i - \pi_i)^2\right] = E\left(z_i^2\right) + E(\varepsilon_i^2)$. Using Lemma 2 and the definition of conesnsus forecasts, we have $MSE^a = E\left[(F\pi_i - \pi_i)^2\right] = E\left(b_i^2\right) + E(\eta_i^2) + E(\varepsilon_i^2) + E(\eta_i \varepsilon_i)$. Under the assumption that η_{ij} and ε_i are uncorrelated, subtracting the previous expressions for MSE^e and MSE^a delivers the desired result. $\square$

Proof of Proposition 1. To start, note that the first part of Assumption 1 and Lemma 2 imply

$$F_j\pi_i = g_i + \alpha z_i + \eta_{ij}, \quad g_i \equiv E\left(F\pi_i \mid X_i\right).$$

Taking averages across forecasters to arrive at consensus forecasts, we obtain

$$F\pi_i = g_i + \alpha z_i + \eta_i, \quad \eta_i \equiv \frac{1}{J_i}\sum_j \eta_{ij}.$$

Next, we can derive the noise in consensus forecasts:

$$\begin{aligned}
E\left(\eta_i^2\right) &= E\left[\left(\frac{1}{J_i}\sum_j \eta_{ij}\right)^2\right] \\
&= E\left[\frac{1}{J_i^2}\left[E\left(\sum_j \eta_{ij}\right)^2 \mid J_i\right]\right] \\
&= E\left[\frac{1}{J_i^2}E\left[\sum_j \eta_{ij}^2 + 2\sum_{j<k}\eta_{ij}\eta_{ik} \mid J_i\right]\right] \\
&= E\left[\frac{1}{J_i^2}\sum_j E\left(\eta_{ij}^2 \mid J_i\right)\right] \\
&= E\left[\frac{1}{J_i}\mathrm{var}\left(\eta_{ij}^2\right)\right] = E\left(\frac{1}{J_i}\right)\Sigma.
\end{aligned}$$

The first equality follows by definition, the second from the law of iterated expectations, the fourth by the third part of Assumption 1, and the fifth by the fourth part of Assumption 1. We can similarly derive the bias in consensus forecasts:

$$\begin{aligned}
E\left(b_i^2\right) &= E\left[\left(g_i + \alpha z_i - x_i - z_i\right)^2\right] \\
&= E\left[\left(g_i - x_i\right)^2\right] + (1-\alpha)^2 E\left(z_i^2\right) \\
&= \Delta + (1-\alpha)^2\Theta.
\end{aligned}$$

Combining the previous two results with Lemma 3 delivers the desired result. □

Proof of Proposition 2. The first part of Assumption 1 and Lemma 2 imply

$$F_j\pi_i = g_i + \alpha z_i + \eta_{ij}, \quad g_i \equiv E\left(F\pi_i \mid X_i\right),$$

which gives $F_{ij}^* = \alpha z_i + \eta_{ij}$. Taking variances and applying the orthgonality condition from Lemma 2 gives

$$\operatorname{var}\left(F_{ij}^{*}\right) = \operatorname{var}(\alpha z_i + \eta_{ij}) = \alpha^2\Theta + \Sigma,$$

delivering the second equation in the proposition. The third equation follows from the third part of Assumption 1 :

$$\operatorname{cov}\left(F_{ij}^{*}, F_{ik}^{*}\right) = \operatorname{cov}\left(\alpha z_i + \eta_{ij}, \alpha z_i + \eta_{ik}\right) = \alpha^2\Theta.$$

Finally, the first equation follows from applying Lemma 1, Lemma 2, and the second part of Assumption 1:

$$\operatorname{cov}(\pi_i^{*}, F_{ij}^{*}) = \operatorname{cov}(z_i + \varepsilon_i, \alpha z_i + \eta_{ij}) = \alpha\Theta.$$

$\square$

Proof of Proposition 3. The first part of Assumption 1 and Lemma 2 imply

$$F_t^j \pi_{it+h} = g_{it}^h + \alpha_h z_{it}^h + \eta_{ijt}^h, \quad g_{it}^h \equiv E\left(F_t \pi_{it+h} \mid X_{it}\right).$$

Then, by definition, $\overline{F}_t^j \pi_{it+h} = g_{it}^h + \alpha_h z_{it}^h$. Forecast revisions are then

$$F_t^j \pi_{it+h} - F_{t-1}^j \pi_{it+h} = \overline{F}_t^j \pi_{it+h} - \overline{F}_{t-1}^j \pi_{it+h} + \eta_{ijt}^h - \eta_{ijt-1}^h.$$

Taking variances, applying the definition of σ_{rev}^2 and $\overline{\sigma}_{rev}^2$, and using the assumption that noise is uncorrelated over time delivers the second equation in the Proposition. Forecast errors are equal to

$$\pi_{it+h} - F_t^j \pi_{it+h} = x_{it}^h + z_{it}^h + \varepsilon_{it}^h - \overline{F}_t^j \pi_{it+h} - \eta_{ijt}^h.$$

The CG coefficient is then

$$\begin{aligned}
\beta_{CG} &= \frac{\operatorname{cov}(\pi_{it+h} - F_t^j \pi_{it+h}, F_t^j \pi_{it+h} - F_{t-1}^j \pi_{it+h})}{\operatorname{var}(F_t^j \pi_{it+h} - F_{t-1}^j \pi_{it+h})} \\
&= \frac{\operatorname{cov}(x_{it}^h + z_{it}^h + \varepsilon_{it}^h - \overline{F}_t^j \pi_{it+h} - \eta_{ijt}^h, \overline{F}_t^j \pi_{it+h} - \overline{F}_{t-1}^j \pi_{it+h} + \eta_{ijt}^h - \eta_{ijt-1}^h)}{\sigma_{rev}^2} \\
&= \frac{\operatorname{cov}(x_{it}^h + z_{it}^h + \varepsilon_{it}^h - \overline{F}_t^j \pi_{it+h}, \overline{F}_t^j \pi_{it+h} - \overline{F}_{t-1}^j \pi_{it+h})}{\sigma_{rev}^2} - \frac{\Sigma_h}{\sigma_{rev}^2} \\
&= \overline{\beta_{CG}}\left(\frac{\overline{\sigma}_{rev}^2 - \Sigma_h}{\sigma_{rev}^2}\right),
\end{aligned}$$

where

$$\overline{\beta_{CG}} \equiv \frac{\text{cov}(x_{it}^h + z_{it}^h + \varepsilon_{it}^h - \overline{F}_t^j \pi_{it+h}, \overline{F}_t^j \pi_{it+h} - \overline{F}_{t-1}^j \pi_{it+h})}{\overline{\sigma}_{rev}^2}.$$

Apply the equation derived for σ_{rev}^2 delivers the result. □

Proof of Proposition 4. From Proposition 1, we have

$$MSE^a = \Delta + (1-\alpha)^2\Theta + \frac{1}{J}\Sigma + E(\varepsilon_i^2).$$

Under the assumption of joint normality,

$$\begin{aligned} F^{e+a}\pi_i &= x_i + E\left(z_i \mid x_i, F\pi_i\right) \\ &= x_i + \frac{\text{cov}(z_i, \alpha z_i + \eta_i)}{\text{var}(\alpha z_i + \eta_i)}\left(F\pi_i - x_i\right) \\ &= x_i + \beta\left[\alpha z + \eta_i\right], \quad \beta = \frac{\alpha\Theta}{\alpha^2\Theta + \frac{1}{J}\Sigma}. \end{aligned}$$

This implies $MSE^{e+a} = E(\varepsilon_i^2) + (1-\beta\alpha)^2\Theta + \beta^2\frac{1}{J}\Sigma$. Subtracting this from MSE^a delivers the result. □

Proof of Proposition 5. First note that combining Assumption 2 with the law of iterated expectations implies

$$\begin{aligned} E\left(\pi_i \mid X_i\right) \equiv x_i &= \left(1-\rho^{h-1}\right)\mu + \rho^{h-1}x_{it}, \\ &\equiv \left(1-\rho^{h-1}\right)\mu + \rho^{h-1}E\left(EPS_{it+1} \mid X_i\right). \end{aligned}$$

Therefore, at horizon h, the bias is

$$\begin{aligned} \Delta &= E\left[\left(E(EPS_{t+h}|X_i) - E(F_t EPS_{t+h}|X_i)\right)^2\right], \\ &= E\left[\left(\rho^{h-1}x_{it} - \rho^{h-1}E(F_t x_{it}|X_i)\right)^2\right] = \rho^{2(h-1)}E\left[\left(x_{it} - E(F_t x_{it}|X_i)\right)^2\right] = \rho^{2(h-1)}\Delta^1. \end{aligned}$$

The noise is

$$\begin{aligned} \Sigma &= var(\eta_{t,h}) = var\left(F_t EPS_{t+h} - E\left(F_t EPS_{t+h}|X_i, Z_i\right)\right), \\ &= var\left(F_t x_{t+h} - E(F_t x_{t+h}|X_i, Z_i)\right) = \rho^{2(h-1)}var(\eta_{t,1}) = \rho^{2(h-1)}\Sigma_\eta^1. \end{aligned}$$

The result follows because $\rho < 1$ by assumption. □

Appendix B. Additional Tables

Table A1. **Variables Included in X_{it}**

This table lists the set of variables that we include in X_{it}, which we use to form our econometric forecast. As described in Section 1.2, we include two lags of each variable. See Appendix C for a detailed discussion of how we use these variables.

Panel A: Collected from WRDS Financial Ratios
The following financial ratios: `capei`, `be`, `bm`, `evm`, `pe_exi`, `pe_inc`, `ps`, `pcf`, `dpr`, `npm`, `opmbd`, `opmad`, `gpm`, `ptpm`, `cfm`, `roa`, `roe`, `roce`, `aftret_eq`, `aftret_invcapx`, `aftret_equity`, `preret_noa`, `pretret_earnat`, `GProf`, `equity_invcap`, `debt_invcap`, `totdebt_invcap`, `capital_ratio`, `int_totdebt`, `cash_lt`, `invt_act`, `rect_act`, `debt_at`, `debt_ebitda`, `short_debt`, `curr_debt`, `lt_debt`, `profit_lct`, `ocf_lct`, `cash_debt`, `fcf_ocf`, `lt_ppent`, `dltt_be`, `debt_assets`, `debt_capital`, `de_ratio`, `intcov`, `intcov_ratio`, `cash_ratio`, `quick_ratio`, `curr_ratio`, `cash_conversion`, `inv_turn`, `at_turn`, `rect_turn`, `pay_turn`, `sale_invcap`, `sale_equity`, `rd_sale`, `adv_sale`, `accrual`, `ptb`, `divyield`

Panel B: Collected from CRSP
2-digit SIC dummies, return in month prior to fiscal year-end, cumulate return in twelve months prior to fiscal year-end excluding last month, trailing 5 year monthly return volatility (all returns adjusted for delisting), stock price on day of fiscal year-end

Panel C: Collected from Compustat
Natural log of total assets, dummies year of fiscal report

Panel D: Collected from I/B/E/S
π_{it}, π_{it-1}, π_{it-2}, number of distinct analysts that issue forecasts in the 45 days following the release of the prior FY report

Appendix C. Additional Details on Supervised Learning Techniques and Forecast Formation

C.1 Supervised Learning Techniques

This section provides a more detailed description of the supervised learning techniques we explore.

Elastic net. The first estimator we explore is Elastic Net, which is defined as follows for a given set of predictor variables X_{it}.

$$\mathcal{L}(\beta,\alpha_1,\alpha_2) \equiv \sum_i \left[(\pi_i - X_{it}'\beta)^2 \right] + \alpha_1 \|\beta\|_1 + \alpha_2 \|\beta\|_2,$$

$$\hat{\beta}^{\text{Lasso}} \equiv \arg\min_{\beta} \mathcal{L}(\beta,\alpha_1,0), \quad \hat{\beta}^{\text{Ridge}}(\alpha_2) \equiv \arg\min_{\beta} \mathcal{L}(\beta,0,\alpha_2),$$

$$\hat{\beta}^{\text{Elastic net}} \equiv \arg\min_{\beta} \mathcal{L}(\beta,\alpha_1,\alpha_2).$$

In order to choose the hyperparameters α_1 and α_2, we use cross-validation on the training set, detailed in Appendix C.2. Intuitively, cross-validation consists of breaking up the training sample into smaller datasets, fitting models on these smaller datasets, and examining which values of the hyperparameters generate the best performance on the other parts of the training set. Importantly, cross-validation is done entirely using the training set to avoid introducing any look-ahead bias.

Tree-based methods. We also consider two tree-based methods: Random Forests (RF) and Gradient-Boosted Trees (GBT). The building block of tree-based estimators are regression trees, which are nonparametric (unlike penalized linear estimators) regression estimators designed to capture arbitrary non-linearities among the variables in X_{it}.

We first describe regression trees, which are "grown" in sequential steps to approximate a function. The tree begins with an initial node containing all observations. Next, this initial node is split into two nodes: observations with $x_{it} < c$ and $x_{it} \geq c$. To make this split, the econometrician chooses the variable $x_{it} \in X_{it}$ and c to minimize MSE. This process of splitting based on a chosen covariate and value continues using the two new subsamples until a terminal criterion is satisfied (e.g. upper bound on the number of observations in each terminal node or the number of splits). The final regression values are then the averages of the outcome variable across all of the observations remaining in each of the terminal nodes.

The process of growing a regression tree immediately illustrates the potential problem with them: they are likely to overfit (i.e. they have high prediction variance), especially if they get extremely large. Without restrictions on the size of the tree, perfect fit in-sample fit could be achieved by having one observation in each terminal node, but this will perform terribly out-of-sample. To address this tendency to over-fit, many "ensemble" methods have been developed, which combine several decision trees with a form of regularization to make more accurate out-of-sample predictions. The two tree-based methods we consider, RF and GBT, are ensemble methods. The core idea behind RF and GBT is to grow many uncorrelated trees and then average their predictions.

RF is constructed based on the intuition of bootstrapping. On each bootstrapped sample, a regression tree with a stopping criterion on the number of splits L with one adjustment - only a random subset of predictor variables are considered at each split.[16] These two steps are then repeated B times, generating B regression trees. Final predictions from the Random Forest are calculated by averaging predictions across the B regression trees. Averaging across many trees, which have different structures due to the randomness in the subset of predictor variables chosen, is the regularization in this method that limits over-fitting and reduces prediction variance. Similar to the penalized linear estimators, the two hyperparameters, $\{B, L\}$ can be chosen using cross-validation on the training set (see Appendix C.2 for details).

GBT starts by fitting a shallow tree of depth d, and calculating the residuals from this regression tree. Then, a second shallow tree of depth d is fit on the residuals calculated from the first tree.[17] This shallow tree is likely has terrible in-sample fit. To improve its fit, a second shallow tree of depth d is fit on the residuals calculated from the first tree. Predicted values are then formed by adding the predicted values from the two trees, shrinking the predicted values from the latter tree by a factor $\lambda \in (0,1)$ (regularization). This procedure is repeated B times, after which the predicted value will be a combination of the predicted value from the first tree and the predicted values of the $B-1$ trees scaled by λ. The sequential growing of trees on (pseudo-)residuals from the previous trees makes the trees less correlated, which is why averaging over trees limits over-fitting. This method has three hyperparameters, $\{B, d, \lambda\}$, which can be chosen using cross-validation on the

[16] If all variables are considered at each split, the procedure of forming many trees across bootstrapped samples is called bagging (i.e. bootstrap aggregation).

[17] Thinking of this procedure as operating on residuals from the trees conveys most of the intuition for why boosting works, but is a technically incorrect description. Gradient-Boosted Trees are a particular form of boosting, where trees are successively fit on pseudo-residuals instead of residuals. Pseudo-residuals are defined as the gradient of the objective function, evaluated at each data point.

training set (see Appendix C.2 for details).

C.2 Formation of Forecasts

This appendix describes the formation of our econometric and econometrician + analyst forecasts, including details on the implementation of our machine learning estimators. For expositional simplicity, we present the procedure as pseudo-code. Additional details on implementation and cross-validation procedures are described at the end of this section.

Pseudo-code. To generate our econometric forecasts using Elastic Net at time t of π_i, the pseudo-code is as follows. For simplicity, denote X_{it} as the set of variables used in the econometric forecast and $X_{it}^{e+a} = \{X_{it}, F_t\pi_{it+h}\}$ as the set of variables used in the econometrician + analyst forecast. We describe our procedure below for our econometric forecasts, but an analogous procedure is used to form econometrician + analyst forecasts, replacing X_{it} with X_{it}^{e+a}.

1. Start with the above dataset that contains X_{it} and π_i, and $F\pi_i$ for each firm-year
2. Replacing all missing values of variables measured at t with industry-time means and then fill all missing values at $t-1$ with values from t and likewise for $t-2$
3. Create year and 2-digit SIC code dummies
4. Initialize $s = 1995$
    (a) Create a **training** dataset of observations indexed by i,s in the following set: $\{(i,t) : t \in \{s-5,...,s-1\}\}$
    (b) Create a **test** dataset of obserations indexed by i,t in the following set: $\{(i,t) : t = s\}$
    (c) Trim all independent variables in the **training** dataset based on 5 times the interquartile range
    (d) Trim all independent variables in the **test** dataset based on 5 times the interquartile range, with the interquartile range *calculated from the training set*
    (e) Standardize all independent variables in the **training** set to have zero mean and unit variance
    (f) Standardize all independent variables in the **test** based on means and variances *calculated from the training set*

(g) Fit a machine learning estimator that is one of the following on the training set, using cross-validation described at the end of this section:

- Elastic Net
- Random Forest
- Gradient-Boosted Trees

(h) Generate forecasts on the test set. Calculating the MSE of these forecasts yields the MSEs for our three forecasts for year s.

(i) Stop if $s = 2021$, otherwise set $s = s+1$ and continue back to (a)

Cross-validation and implementation details by estimator. We use the following cross-validation and implementation procedures for each machine learning algorithm on our training sets for each model. All procedures are implemented using the `sklearn` package in `Python 3.9`. We use default inputs to all `sklearn` functions mentioned below, unless otherwise specified.

- Elastic Net: We use 5-fold cross-validation on the training set, implemented using the `ElasticNetCV` function in `sklearn`. We search over a grid of the parameter `l1_ratio` $\in [0.1, 0.99]$, which corresponds to the ratio of the $\mathcal{L}^1$ to $\mathcal{L}^2$ penalty parameters.

- Random Forest: We use 5-fold cross-validation on the training set, implemented using the `GridSearchCV` function for `RandomForestRegressor` in `sklearn`. We set `n_estimators` to 1000, corresponding to the number of decision trees in the ensemble, and search over the following grid for each parameter: `max_depth` $\in [4, 8]$, `max_features` $\in [0.3, 1]$, `min_samples_leaf` $\in [1, 5]$, and `min_samples_splits` $\in [2, 10]$. We use bootstrap samples for each decision tree. These parameter choices are similar to Gu et al. (2018) and Hansen and Thimsen (2020).

- Gradient-Boosted Trees: We use 5-fold cross-validation on the training set, implemented using the `GridSearchCV` function for `GradientBoostingRegressor` in `sklearn`. We search over the following grid for each parameter: `n_estimators` $\in [500, 10000]$, `max_depth` $\in [1, 3]$, and `learning_rate` $\in [0.001, 0.1]$. These parameter choices are similar to Gu et al. (2018).

Hardware. Rolling estimation with repeated cross-validation is computationally-intensive. We parallelize each model estimation across 96 CPUs on the MIT SuperCloud server, with each estimation taking around 100 days of CPU time.

Appendix D. Interpretive Model for MSS Normalization

We write down here a simple model that gives a simple interpretation to the normalized MSE that we use throughout the paper. Take the perspective of a hypothetical agent, who seeks to allocate capital across firms. We assume that investing k_i dollars in firm i eventually generates cash-flows $\pi_i k_i - \frac{1}{2\gamma} k_i^2$. γ is a measure of returns to scale ($\gamma = \infty$ corresponds to constant returns to scale). This agent is risk neutral and therefore maximizes the sum of all expected cash-flows:

$$\Pi_h = \sum_i \left(k_i F \pi_i - \frac{1}{2\gamma} k_i^2 \right)$$

where the expectation is taken using the agent's forecasting rule F. In this simple problem, capital allocation for firm i is $k_i = \gamma F \pi_i$. We compare this allocation to the perfect foresight allocation $k_i^{PF} = \gamma \pi_i$. Trivially, the perfect foresight allocation dominates all forecast-based allocations (including rational ones).

The expected cash-flow loss relative to perfect foresight allocation can then be written as:

$$\frac{\Pi_h^{PF} - \Pi_h^F}{\Pi_h^{PF}} = \frac{MSE_h^F}{MSS_h} \tag{16}$$

where MSE_h^F is the MSE of the forecasting rule and MSS_h is the realized mean of π_i^2. Thus, (16) shows normalizing mean-squared errors by the mean squared EPS can be interpreted as the percent allocative loss compared to a perfect foresight optimizer.

Appendix E. Additional Details on GMM Estimation

In this appendix, we discuss the details of our GMM estimation based on the following moment conditions from Proposition 2:

$$
\begin{aligned}
E\left(\pi_i^* F_{ij}^*\right) &= \alpha\Theta, \\
E\left[\left(F_{ij}^*\right)^2\right] &= \alpha^2\Theta + \Sigma, \\
E\left(F_{ij}^* F_{ik}^* \mid j \neq k\right) &= \alpha^2\Theta.
\end{aligned}
$$

We have now replaced covariances with second moments given all variables are mean zero. The computation of these expectations requires further clarification, given the third moment varies at a different level than the first two.

We start with an i-t-j panel of individual analyst forecasts for each firm-year discussed in Section 1.1. Denote the size of this dataset as N_0. We then compute all possible interactions between the forecasts of the J_{it} analysts following each firm, resulting in $\binom{J_{it}}{2}$ interactions per i-t. We use this set of interactions for each i-t to construct an i-t-j-k panel that contains the forecasts of analysts j and k and their interaction for all possible j-k pairs. Denote the size of this dataset as N_1. We denote sample expectations taken on this i-t-j-k panel as $\widehat{E}$.

The score vector we use in our GMM estimation (making t explicit for clarity):

$$
m(\pi_{it}^*, F_{itj}^*, F_{itk}^*; \alpha, \Theta, \Sigma) = \begin{pmatrix} \frac{N_1}{N_0}\pi_{it}^* F_{itj}^* - \alpha\Theta \\ 1(j=k)\frac{N_1}{N_0}F_{itj}^* F_{itk}^* - \alpha^2\Theta - \Sigma \\ 1(j \neq k)\frac{N_1}{N_1 - N_0}F_{itj}^* F_{itk}^* - \alpha^2\Theta \end{pmatrix}.
$$

Note there is reweighting based on N_1 and N_0. This reweighting ensures that taking the expectation of this score vector on an i-t-j-k panel delivers the same result that we would get by calculating these moments on an i-t-j panel. However, we cannot use an i-t-j panel for an estimation because performing GMM requires sample expectations of each moment condition to be calculated as sample averages on the same dataset.

Our final step is to generate our parameter estimates by solving:

$$
\left(\hat{\alpha}, \hat{\Theta}, \hat{\Sigma}\right) = \underset{\alpha,\Theta,\Sigma}{\arg\min}\, \widehat{E}\left[m(\pi_i^*, F_{itj}^*, F_{itk}^*; \alpha, \Theta, \Sigma)\right]' \widehat{E}\left[m(\pi_i^*, F_{itj}^*, F_{itk}^*; \alpha, \Theta, \Sigma)\right].
$$

We solve this optimization problem using a global basinhopping algorithm implement in `SciPy`.

As we did with our MSEs, we normalize by the average squared EPS (calculated using an unweighted average on the i-t-j panel). We perform this estimation procedure separately for each h.

To calculate standard errors, we calculate the Jacobian of $m(\pi^*_{it}, F^*_{itj}, F^*_{itk}; \alpha, \Theta, \Sigma)$ with respect to our three parameters analytically. We then use the standard GMM formula for standard errors after estimating the covariance matrix of residuals in the usual way. For Δ, we compute the standard error simply as the standard deviation divided by $\sqrt{N_0}$, since it's just a mean. These standard errors are likely to be too tight because they ignore sampling uncertainty in the predictions generated from our machine learning estimators, effectively treating them as raw data that is resampled directly to compute moments. Here we are constrained by the lack of asymptotic results that characterize the behavior of our statistical learning estimators in large samples. This is approach is standard in the literature (e.g., Patton and Verardo 2012).[18]

Weighting in Table 3 MSE decomposition. Because of the weighting in our GMM, our resulting parameter estimates equally-weight all firm-year-analyst observations (i.e. firms with more analysts will be weighted more). However, the MSEs reported in Table 2 and Table 4 equally weight all firm-year observations because they only vary at the firm-year level. To address this issue, we recompute the expectations required to calculate MSE^a, MSE^a, and $\frac{1}{J}$ on the i-t-j panel used for GMM estimation when reporting their values in Table 3.

[18] We could in principle bootstrap the data and re-estimate our ML models, but this is too computationally intensive. Moreover, it's not clear that every estimator satisfies the regularity conditions required for the bootstrap to be asymptotically valid (Horowitz 2001).

Appendix F. Afrouzi et al. (2021) Model

In this section, we apply the model proposed by Afrouzi et al. (2021) to our setting. We first briefly describe it, referring the reader to Afrouzi et al. (2021) Section 5 for additional details. We then show it qualitatively delivers an upward sloping term structure of bias and noise, but fails quantitatively in our setting for an intuitive reason.

F.1 Model Description

Consider the same setup as in Section 4.1, including Assumption 2. In addition, assume further that $u_{it}^x \sim \mathcal{N}(0, \sigma_u^2)$ and allow for the mean of x_{it} to be different from zero, denoted μ. At time t, the analyst needs to form forecasts of EPS_{it+h} for $h \geq 1$. The analyst costlessly observes x_{it-1}, z_{it}^h, and ρ. Importantly, the analyst does not know μ, but has the prior $\mu | x_{it-1} \sim \mathcal{N}(x_{it-1}, \underline{\tau})$. The analyst also knows the DGP for EPS_{it}. Because x_{it}^h and z_{it}^h are orthogonal, we assume the analyst forms forecasts of EPS_{it+h} by first forming optimal forecasts of x_{it+h-1} and z_{it}^h individually, then adding them together. Denote her forecasts by F_t. We assume $\alpha = 1$ for simplicity to focus on forecasts of x_{it}, which we show fail to match the data.

To form $F_t x_{it+h}$, the analyst has access to two sources of information. First, the agent costlessly observes x_{it-1}, which is "on top of his mind". Secondly, the agent can retrieve additional information from past data to learn about μ, but doing so is costly. The chooses the level of information acquisition in order to minimize the expected squared error of his forecasts, subject to the cost of information retrieval. Formally, the analyst solves the following problem:

$$\min_{S_{it}} E\left[\min_{F_t x_{it+h}} E\left[(F_t x_{it+h} - x_{it+h})^2 | S_{it}\right] + C(S_{it})\right], \tag{17}$$
$$s.t. \quad \{x_{it-1}\} \in S_{it} \subset \mathcal{S}_{it}, \quad \mathcal{S}_{it} = \{s : s \perp \mu | x_{it-1}, x_{it-2}, \ldots\}.$$

The solution to the inner maximization problem is $F_t x_{it+h} = E(x_{it+h} | S_{it})$. As shown in Afrouzi et al. (2021), the assumption that u_{it} is normally distributed implies the outer maximization problem can be reduced to choosing a belief prediction about μ, denoted τ. Denote the solution to this problem as $\tau^*(h)$.

To further characterize the solution to this problem, Afrouzi et al. (2021) assume the cost of information retrieval takes the following form:

$$C_t(S_t) \equiv \omega \frac{\exp(2\ln(2)\cdot\gamma\cdot\mathbb{I}(S_t,\mu\,|\,x_{it-1}))-1}{\gamma}.$$

In this expression, $\mathbb{I}(\cdot,\cdot)$ denotes Shannon mutual information, $\omega \geq 0$ governs the overall cost of retrieval, and $\gamma \geq 0$ measures convexity of the cost function in mutual information. Given this cost function, Afrouzi et al. (2021) show the choice of belief precision, τ, that solves (17) is:

$$\tau^*(h) = \underline{\tau}\max\left\{1, \left(\frac{\left(1-\rho^{h+1}\right)^2}{\omega\underline{\tau}}\right)^{\frac{1}{1+\gamma}}\right\}. \tag{18}$$

This choice of $\tau = \tau^*(h)$ implies the analyst's forecast is the following:

$$F_t x_{it+h} = (1-\rho^{h+1})\left(1-\frac{\underline{\tau}}{\tau^*(h)}\right)\mu + \left(\rho^{h+1}+(1-\rho^{h+1})\frac{\underline{\tau}}{\tau^*(h)}\right)x_{it-1} + \eta^h_{it},$$
$$\eta^h_{it} \sim \mathcal{N}\left(0, (1-\rho^{h+1})^2\frac{1}{\tau^*(h)}\left(1-\frac{\underline{\tau}}{\tau^*(h)}\right)\right).$$

Note that $F_t x_{it+h}$ is a random variable because of the expectation noise, η^h_{it}.

F.2 Term Structure of Bias and Noise

Now that we have characterized the analyst's forecasts, we can determine the term structure of bias and noise in this model.

Proposition F1. *In the Afrouzi et al. (2021) model, the term structure of bias and noise are given by:*

$$\Delta_h = \left[(1-\rho^h)\frac{\underline{\tau}}{\tau^*(h-1)}\right]^2\frac{\sigma_u^2}{1-\rho^2}, \qquad \Sigma_h = \left(1-\rho^h\right)^2\frac{1}{\tau^*(h-1)}\left(1-\frac{\underline{\tau}}{\tau^*(h-1)}\right).$$

Thus, the Afrouzi et al. (2021) model produces an upward sloping term structure of bias and noise.

Proposition F1 shows noise in increasing with the horizon because of $\tau^*(h-1)$ increasing with the horizon (see eq. (18)). The intuition here is that at longer horizons, it is more useful to know the long-run mean, so the analyst engages in more information retrieval. This greater retrieval provides the analyst with more noisy information, creating large noise in her forecasts.

Additionally, the formula for Δ_h in Proposition F1 shows bias is increasing with the horizon, since $\tau^*(h-1)$ is increasing in h at a slower rate than $1-\rho^h$. This upward term structure of bias results from a combination of two effects. On the one hand, the agent engages in more retrieval, moving her forecast closer to the conditional expectation. On the other hand, the analyst still overweights x_{it-1} in her estimate of the long-run mean, and this overweighting is magnified at longer horizons because the analyst's forecast moves closer to her subjective expectation of the long-run mean. This second effect dominates, generating an upward sloping term structure of bias.

In sum, the Afrouzi et al. (2021) model generates an upward sloping term structure of bias and noise. We now discuss how this model cannot *quantitatively* fit the data.

F.3 Quantitative Model Fit

The Afrouzi et al. (2021) model has 5 parameters: ρ, σ_u, ω, γ, and $\underline{\tau}$. The key endogenous parameter is $\tau^*(h)$. Rewriting the equation for Δ_h in Proposition F1, we obtain an expression for $\tau^*(h)$ as a function of the bias at horizon h:

$$\frac{\tau^*(h-1)}{\underline{\tau}} = \frac{\left(1-\rho^h\right)\sigma_u}{\sqrt{\Delta_h\left(1-\rho^2\right)}}. \tag{19}$$

Equation (19) is useful because it express $\frac{\tau^*(h-1)}{\underline{\tau}}$, which is the crucial to this model, as a function of data moments we have already estimated (or can estimate).

We focus on annual forecast horizons and estimate $\rho = 0.862$ and $\sigma_u = 0.00857$ by regressing $F_t\pi_{it+2}$ onto $F_t\pi_{it+1}$. Using our values of Δ_h from Figure 2, we can evaluate (19) at each horizon:

$$\frac{\left(1-\rho^1\right)\sigma_u}{\sqrt{\Delta_1\left(1-\rho^2\right)}} \approx 0.28, \quad \frac{\left(1-\rho^2\right)\sigma_u}{\sqrt{\Delta_2\left(1-\rho^2\right)}} \approx 0.26, \quad \frac{\left(1-\rho^3\right)\sigma_u}{\sqrt{\Delta_3\left(1-\rho^2\right)}} \approx 0.26.$$

By eq. (19), we obtain

$$\frac{\tau^*(1-1)}{\underline{\tau}} \approx 0.28, \quad \frac{\tau^*(2-1)}{\underline{\tau}} \approx 0.26, \frac{\tau^*(3-1)}{\underline{\tau}} \approx 0.26.$$

However, this contradicts eq. (18), which shows $\forall h$, $\frac{\tau^*(h)}{\underline{\tau}} \geq 1$. Thus, these numerical results illustrate that the Afrouzi et al. (2021) model cannot match our data because it cannot generate public bias.

The intuition of the quantitative failure of the Afrouzi et al. (2021) model to match our data is the following. From Proposition F1, the public bias at horizon h is increasing in σ_u: as $\sigma_u \to 0$, there will be no public bias because x_{it} will be a deterministic process. In the data, we in fact find σ_u is quite low. However, despite a low σ_u, we still find substantial forecasting bias – the public bias is so high that this model would require the analyst to *forget* information to match this level of bias. The heart of this problem is that the model gives the analyst access to x_{it-1}, which is really close to x_{it+h} because σ_u is low. In other words, the fact that the analyst gets to see the machine forecast from the last period, $F^m_{t-1}EPS_{it} = x_{it-1}$, gives her too much knowledge to be as biased (and noisy) as we find in the data.