Università Commerciale "Luigi Bocconi"

PhD School

PhD program in: Economics and Finance

Cycle: XXXV

Disciplinary Field (code): SECS-P/05

Essays in Finance

Advisor: Carlo Ambrogio Favero

PhD Thesis by

Gabriele Confalonieri

ID number: 3105995

Year: 2024

Contents

Acknowledgements

I thank my mentors Carlo Ambrogio Favero, Massimiliano Marcellino and Michael McMahon for their guidance and support. I thank Fabio Bagliano and Andrea Tamoni for their precious feedback on my research work. I thank all the professors I worked with during my PhD, in particular Fulvio Ortu, Anna Battauz, Andrea Sironi, Andrea Resti, Giampaolo Gabbi, Carlo Cottarelli, Nicola Gennaioli, and Thorsten Martin. I thank all my PhD colleagues, in particular Giuliano Graziani and Alireza Aghaee Shahrbabaki.

I thank my wife Elizaveta, my family and friends for their support.

Introduction

In the first chapter of this PhD Thesis, we investigate why large cross-sections of long-short anomaly portfolios predict the market excess return. We develop an econometric model for the prices of the long and short legs of the anomalies. Using dimension reduction techniques, we show that their deviations from equilibrium predict the aggregate market return. This result holds at multiple horizons and is mostly driven by the long components of the anomaly portfolios. We interpret these findings through an asymmetric limits of arbitrage model with slow-moving capital.

In the second chapter, we compare the information contained in the headlines and the full text of more than $400,000$ business news articles. We show that sentiment measures extracted from the two sources are highly correlated. Using state-of-the-art machine learning methods, headline-based forecasts of macroeconomic indicators have equal or greater accuracy out-of-sample than forecasts based on the whole text. We interpret our findings through a model of news with attention costs and beauty contest elements.

In the third chapter, we investigate whether measures of sentiment extracted from quarterly earnings conference-calls affect the dynamics of stock prices. Using a cross-section of publicly traded companies, we show that sentiment positively correlates with price deviations from their long-run trend, estimated via an error correction model. We document that even though sentiment does not predict future stock returns, it impacts the speed at which prices revert to equilibrium. We find asymmetric effects on overpriced and underpriced stocks.

Chapter 1

Price Trends and Return Predictability

1.1 Introduction

Recent empirical evidence shows that large cross-sections of long-short anomaly portfolios predict the market excess return (Dong et al., 2022). This chapter investigates why. I show that the *price* deviations of the long and short legs of the anomaly portfolios predict the market excess *return*. This result holds at multiple horizons and it is mostly driven by the long components of the anomalies. I interpret these findings through an asymmetric limits of arbitrage model with slow-moving capital.

First, I develop an econometric model for the prices of the long and short components of anomaly portfolios. I consider 100 anomalies from the cross-sectional asset pricing literature. They are long-short portfolios built by sorting stocks into deciles based on a given characteristic (e.g. the book-to-market ratio), and going long (short) the tenth (first) decile. I consider the long and short legs of each portfolio separately and, in the spirit of Fama and French (1988), I decompose their prices into a permanent and a transitory component. The former is common across assets. Following Bai (2004) and Banerjee and Marcellino (2009), I estimate it via cross-sectional principal components of prices. The transitory component is the asset-specific deviation of the prices from the common trend.

This permanent-transitory decomposition gives rise to an error correction mechanism, through which current price deviations predict future returns. Indeed, suppose that the price of a given asset is above the equilibrium. Its returns in the following periods will necessarily be lower, so that the price can move back towards the trend-implied level and the temporary mispricing is corrected.[1] Favero et al. (2021) find evidence for this error correction mechanism in a large cross-section of equity portfolios. The authors show that such predictability channel holds at the individual asset level, that is the mispricings of a given asset predict its own future return. In this chapter, I document that this mechanism can be used to predict the return of the aggregate market as a whole from the individual assets.

After estimating the common stochastic trends and the error correction terms for each portfolio leg, I use them to predict the market returns. I regress the return of the market between time t and $t+h$ on all the error correction terms at time t. Since the number of regressors is potentially large (2 legs $\times$ 100 portfolios), I use dimension reduction methods that guard against overfitting. In particular, I use shrinkage techniques such as principal components and Elastic Net. I compare my results with Dong et al. (2022), using the (lagged) returns of each anomaly leg as regressors. I evaluate the models through out-of-sample tests, since they are the most relevant evidence for stock return predictability (Welch and Goyal, 2008; Martin and Nagel, 2021).

I find that the mispricings of the long and short anomaly legs predict the market excess return one month in the future. The principal component method delivers an out-of-sample R^2 (R^2_{OS}) of 0.56%, which is above the 0.5% threshold for economic significance by Campbell and Thompson (2008). This figure is also statistically significant, according to the Clark and West (2007) test. I run the out-of-sample exercise on the sub-samples of long and short legs to disentangle their respective contribution. The R^2_{OS} for the long

[1] I consider a framework in which deviations from the long-run equilibrium reflect mispricing, as in Dong et al. (2022). While I do not consider a risk-based environment, this could be an interesting avenue for future research.

portfolios ranges from 0.74% to 1.41%, both economically and statistically significant. Instead, the short legs deliver insignificant figures. Repeating the same exercise on the anomaly returns, rather than the error correction terms, gives an opposite picture. When using both long and short portfolio returns, the R^2_{OS} is either negative or statistically insignificant. As in Dong et al. (2022), the short leg returns forecast the market, while the out-of-sample predictive ability of the long legs is weak.

I repeat the same out-of-sample exercise at longer horizons and the results are even stronger. The error correction terms of both anomaly legs strongly predict the market one quarter and one year in the future. These results are mostly driven by the strong R^2_{OS} of the long legs, which ranges from 1.38% to 2.07% at the quarterly horizon, and from 2.81% to 13.10% at the annual one. These figures are both economically and statistically significant. On the contrary, the out-of-sample performance of the short legs remains weak. These findings are in stark contrast with Dong et al. (2022), since the anomaly returns do not predict the market at horizons longer than one month. This fact shows the importance of modeling prices directly in the empirical analysis, as the information on long-run relations is differenced away in the cross-section of returns.

As a robustness test, I conduct an asset allocation analysis. I consider a risk-averse investor who can invest in the equity market and a risk-free asset. I compare the portfolios built using the error correction terms vis-a-vis the historical average to forecast the market excess return. The performance of the two portfolios is ranked in terms of the utility to the investor, which is a measure of the economic value of the forecasting strategies. In line with my previous results, the mispricings of the long legs lead to sizable utility gains, that range from 1.22% to 6.38% and increase with the investment horizon. The opposite holds for the short legs, with utility gains from -2.72% to 0.14%.

I interpret my findings within an econometric framework that combines the permanent-transitory price decomposition (Fama and French, 1988) with asymmetric limits of arbitrage (Shleifer and Vishny, 1997). I consider a data generating process in which the short (long) legs are characterized by more (less) persistent over-pricing (under-pricing) shocks. This assumption is consistent with the fact that *i*) short legs are over-priced, so that they deliver lower returns and the anomaly portfolios generate "alphas"; *ii*) arbitrage with respect to overpriced shares is less aggressive, e.g. because of short-selling constraints (Miller, 1977); *iii*) in the data, the persistence of the short legs' error correction terms is larger. In this econometric framework I prove that, ceteris paribus, the more persistent mispricings of the short legs have lower predictive power. (Intuitively, in the limit of permanent price deviations, the short legs would be useless for forecasting.) I also show that modeling directly price deviations, rather than the bare returns, improves the forecasting performance, especially at longer horizons.

Finally, I provide a microfoundation to the data generating process that underlies my analysis. I develop a model with slow-moving capital (Duffie, 2010), in which a mass of inattentive investors can only trade once every k periods. I show that such a model can account for temporary deviations of prices from their fundamental value, which are slowly corrected towards the equilibrium. I study the implications of asymmetric limits of arbitrage for the price dynamics in such a setting, introducing short selling costs (Gromb and Vayanos, 2010). As posited in the econometric framework, this financial friction significantly affects the persistence of price deviations.

Related Literature This chapter is related to Dong et al. (2022), which studies the link between cross-sectional and time series stock return predictability. By using dimension reduction techniques, the authors show that *i*) a large number of anomaly returns predict the market excess return one month ahead; *ii*) this result is mostly driven by the short leg returns; *iii*) the predictability vanishes at longer horizons. They argue that these findings are consistent with a data generating process with stationary components in the prices of the long and short legs of the anomaly portfolios. In this chapter, I bring their theoretical framework to the data and I focus on anomaly prices rather than returns. I predict the market at horizons up to one year – a key advantage of estimating long run relations in the cross-section of prices. Moreover, recent contributions by Cakici et al. (2024) challenge the approach followed by Dong et al. (2022), claiming that anomalies cannot predict aggregate market returns. According to their analysis, any predictability cannot be extended beyond the US stock market and it is driven by groups of specific anomalies. Even though my analysis still focuses on the US stock market (while the latter authors use an international dataset), I show that using prices rather than returns reconciles this gap, improving the forecasting performance, and there is evidence of robustness.

This work is also related to a literature that models both asset returns and prices, which starts with the seminal contribution by Fama and French (1988). Recent studies in this area include Baba Yara et al. (2020), van Binsbergen et al. (2021), Cho and Polk (2020), Cohen et al. (2009). In particular, Favero et al. (2021) perform a permanent-transitory decomposition of anomaly portfolio prices and study its implications for return predictability. The authors focus on predicting the return of each anomaly separately. Instead, this chapter uses the deviations from equilibrium of the anomaly portfolios to forecast the aggregate market excess return. I also propose a new way to estimate the common stochastic trends in the cross-section of test assets.

This chapter relates to the literature of non-stationary factor models. I use the methods introduced by Bai (2004), Banerjee and Marcellino (2009) and Banerjee et al. (2014) to estimate the common stochastic trends in the cross-section of prices. This chapter brings their insights to the equity market. Differently from these studies, I use the deviations from equilibrium of each series to predict the market excess return.

1.2 The Econometric Framework

In this section I describe the econometric framework that guides my empirical analysis. The model is based on Dong et al. (2022). I derive some predictions that I will later test in the data. I consider a data generating process that will be microfounded in Section 1.6, in a model with slow-moving capital.

1.2.1 The Model

Consider a securities market in which the assets are sorted in two portfolios according to an observable characteristics X. The Long (Short) portfolio, indexed by $l = L$ ($l = S$), contains the assets with the highest (lowest) value of X. As an example, if X were the book-to-market ratio, the L and S portfolios would correspond to the value and growth stocks respectively. I denote the log-price of each portfolio at time t by $\log P_{l,t}$, so that its h-period log-return (cum-dividend) is $r_{l,t,t+h} = \log P_{l,t+h} - \log P_{l,t}$.

In the spirit of Fama and French (1988), I assume that the log-prices of the portfolios are driven by a permanent and a transitory component, denoted with $\log F_{l,t}$ and $u_{l,t}$ respectively. That is

$$\log P_{l,t} = \log F_{l,t} + u_{l,t}, \tag{1.1}$$

$$u_{l,t} = \rho_l u_{l,t-1} + \eta_{l,t}, \tag{1.2}$$

$$\log F_{l,t} = \log F_{l,t-1} + \nu_{l,t}, \tag{1.3}$$

with $\rho_l < 1$ for $l = L, S$. In contrast with Dong et al. (2022), the two legs are driven by two different trends, i.e. $\log F_{L,t} \neq \log F_{S,t}$. This latter assumption implies that the price of the anomaly long-short portfolio $\log P_{L,t} - \log P_{S,t}$ is non stationary, which is consistent with recent empirical evidence on the cross-section of asset prices (Favero et al., 2021).

By taking first differences of Equation (1.1), we have

$$r_{l,t,t+1} = f_{l,t,t+1} + (\rho_l - 1)\, u_{l,t} + \eta_{l,t+1}, \tag{1.4}$$

with $f_{l,t,t+1} := \log F_{l,t+1} - \log F_{l,t}$. If $\rho_l < 1$, i.e. if there is cointegration, then $u_{l,t}$ is an error correction term (ECT, henceforth) which implies that prices move towards the equilibrium. As an example, suppose that $u_{l,t} > 0$, i.e. the price of portfolio l at time t is above the equilibrium level implied by the trend. Since $(\rho_l - 1) < 0$, the return from t to $t+1$ will be lower, and $\log P_{l,t+1}$ will move downwards to correct the mispricing. Another key implication is that $u_{l,t}$ is a predictive term in Equation (1.4): it is estimated at time t, and it provides information on the return between time t and $t+1$.

So far I have described an error correction mechanism that drives the dynamics of prices, and that can be exploited to predict the returns of *each* portfolio. Now I will explore its implications for the *aggregate* portfolio, i.e. the market. In the context of my simplified framework, the market portfolio is an equally weighted portfolio of L and S. Its return between time t and $t+1$ is

$$r_{M,t,t+1} = \frac{1}{2} r_{L,t,t+1} + \frac{1}{2} r_{S,t,t+1} \tag{1.5}$$

$$= f_{t,t+1} + \frac{1}{2}(\rho_L - 1) u_{L,t} + \frac{1}{2}(\rho_S - 1) u_{S,t} + \eta_{t+1}, \tag{1.6}$$

with $f_{t,t+1} = 1/2 f_{L,t,t+1} + 1/2 f_{S,t,t+1}$ and $\eta_{t+1} = 1/2\eta_{L,t+1} + 1/2\eta_{S,t+1}$. Equation (1.6) is obtained by substituting $r_{L,t,t+1}$ and $r_{S,t,t+1}$ with their expressions from Equation (1.4). Interestingly, Equation (1.6) shows that the return predictability at the individual portfolio level has implications for the overall market. The deviations from equilibrium of the two portfolios, $u_{L,t}$ and $u_{S,t}$, predict $r_{M,t,t+1}$, as long as $\rho_L, \rho_S < 1$.

The same line of reasoning can be applied to compute h-period returns. Starting from Equation (1.1) the return between time t and $t+h$ of portfolio l is

$$r_{l,t,t+h} = f_{l,t,t+h} + \left(\rho_l^h - 1\right) u_{l,t} + \sum_{i=0}^{h-1} \rho^i \eta_{l,t+h-i}, \tag{1.7}$$

with $f_{l,t,+h} := \log F_{l,t+h} - \log F_{l,t}$. Equation (1.7) is in stark contrast with Dong et al. (2022), who argue that the predictability channel due to deviations from equilibrium is short lived. Due to their persistence ρ_l, the mispricings are not necessarily corrected in one period. This fact implies that they hold information to predict the market at longer horizons. The same insights apply to the market portfolio, whose h-period return is

$$r_{M,t,t+h} = \frac{1}{2} r_{L,t,t+h} + \frac{1}{2} r_{S,t,t+h} \tag{1.8}$$

$$= f_{t,t+h} + \frac{1}{2}\left(\rho_L^h - 1\right) u_{L,t} + \frac{1}{2}\left(\rho_S^h - 1\right) u_{S,t} + \eta_{t+h}, \tag{1.9}$$

with $f_{t,t+h} = 1/2 f_{L,t,t+h} + 1/2 f_{S,t,t+h}$ and $\eta_{t+h} = 1/2\eta_{L,t+h} + 1/2\eta_{S,t+h}$. Notice that the error term in equation (1.9) is a sum of error terms over multiple horizons, as the last term in Equation (1.7). Equations (1.7) and (1.9) highlight the importance of modeling both prices and returns. Empirical analyses that only focus on the latter omit relevant information to predict at longer horizons.

The persistence of mispricing ρ_l of the two portfolios is key in this framework. Indeed, the (cointegration) condition $\rho_l < 1$ implies that the residuals $u_{l,t}$ predict future market returns. ρ_l is related to the speed with which the price of portfolio l goes back to the equilibrium level. It is therefore natural to differentiate the Long and the Short portfolios in terms of their mispricing persistence. In particular, I assume that $\rho_L < \rho_S$, which is in line with asymmetric limits of arbitrage.

Assumption 1 (Asymmetric Limits of Arbitrage)**.** *The transitory component of the Short portfolio price ($u_{S,t}$) is more persistent than the one of the Long portfolio price ($u_{L,t}$). That is, $\rho_L < \rho_S < 1$.*

To understand the intuition behind Assumption 1, recall that the long and short portfolios in the theoretical framework mirror the long and short legs of the anomaly portfolios from the cross-sectional literature. That is, the long legs are expected to generate relatively high returns, and they are considered to be underpriced. The opposite holds for the short legs, so that the long-short anomaly portfolios generate "alphas". Under asymmetric limits of arbitrage, it is "easier" to correct underpricing rather than overpricing (Shleifer and Vishny, 1997). As an example, short-sale constraints should make overpricing more prevalent and persistent than underpricing (Miller, 1977). Imposing $\rho_L < \rho_S$ captures this financial friction in reduced form.

1.2.2 The Predictions

Suppose we have data on the prices of the two portfolios and we want to predict the market returns. Knowing that the data generating process is as in Equation (1.1), we run regressions in the form

$$r_{M,t,t+1} = \alpha_l + \beta_l u_{l,t} + \epsilon_{l,t+1}, \quad l = L, S. \tag{1.10}$$

The coefficient β_l and the mean squared forecast error are the objects of interest. Comparing Equation (1.10) with the expression for the market returns in Equation (1.6), I derive the following proposition.

Proposition 1 (Long vs Short Legs)**.** *Under Assumption 1, the ratio between the Long and Short portfolios standardized regression coefficients is*

$$\left| \frac{\tilde{\beta}_L}{\tilde{\beta}_S} \right| = \sqrt{\frac{1-\rho_L}{1-\rho_S}} \sqrt{\frac{1+\rho_S}{1+\rho_L}} > 1. \tag{1.11}$$

Moreover, the ratio between the mean squared forecast errors is

$$\frac{\mathbb{E}\left[\epsilon^2_{L,t+1}\right]}{\mathbb{E}\left[\epsilon^2_{S,t+1}\right]} < 1. \tag{1.12}$$

Proof. See Appendix A. □

According to Proposition 1, the ECT of the Long portfolio is a better predictor than the ECT of the Short one: the former delivers a larger (in absolute value) standardized β and lower expected forecast errors. To see why this result holds in an intuitive way, consider the limit $\rho_S \to 1$. Equation (1.6) shows that the coefficient on $u_{S,t}$ would be zero. Indeed, in this case any mispricing shock to the Short portfolio would be permanent, and there is no mechanism to reverse it towards the equilibrium. Hence, the predictability channel would be off.

Consider now an alternative regression model, in which the predictive term is the lagged return of the individual portfolios. That is

$$r_{M,t,t+1} = \alpha_{r,l} + \beta_{r,l} r_{l,t-1,t} + \epsilon_{r,l,t+1}, \quad l = L, S, \tag{1.13}$$

as in Dong et al. (2022). I use Equation (1.4) to obtain an analytical expression for $\beta_{r,l}$, and I compare it to the ECT coefficient β_l in the following proposition.

Proposition 2 (ECT vs Returns)**.** *Under $\rho_l < 1$, the ratio between the standardized regression coefficients is*

$$\left| \frac{\tilde{\beta}_{r,l}}{\tilde{\beta}_l} \right| < 1. \tag{1.14}$$

Moreover,

$$\frac{\mathbb{E}\left[\epsilon^2_{r,l,t+1}\right]}{\mathbb{E}\left[\epsilon^2_{l,t+1}\right]} > 1. \tag{1.15}$$

Proof. See Appendix A. □

Proposition 2 shows that the ECT $u_{l,t}$ is a better predictor than the lagged return $r_{l,t-1,t}$. This fact holds in terms of both the standardized regression coefficients and the mean squared forecast errors, as long as $\rho_l < 1$. The proof of the Proposition shows why $r_{l,t-1,t}$ still predicts the market as in Dong et al. (2022). Intuitively, $r_{l,t-1,t}$ can be considered as an instrument for $u_{l,t}$: it depends on $u_{l,t-1}$ through the error correction mechanism in Equation (1.4), thus containing information on $u_{l,t}$ due to the ECT persistence.

Finally, consider a regression in the form

$$r_{M,t,t+k} = \alpha_l^k + \beta_l^k u_{l,t} + \epsilon_{l,t+k}, \quad l = L, S, \tag{1.16}$$

predicting the k-period ahead market return. Equation (1.9) let us obtain the analytical expression for β_l^k and compare it at different horizons.

Proposition 3 (Multi-horizon Regressions). *Under cointegration ($\rho_l < 1$), the ratio between the standardized regression coefficients is*

$$\left| \frac{\tilde{\beta}_l^{k+n}}{\tilde{\beta}_l^k} \right| > 1, \quad n \geq 1. \tag{1.17}$$

Moreover,

$$\frac{\mathbb{E}\left[\epsilon_{l,t+k+n}^2\right]}{\mathbb{E}\left[\epsilon_{l,t+k}^2\right]} < 1. \tag{1.18}$$

Proof. See Appendix A. □

Proposition 3 shows that the predictive power increases with the forecast horizon. Once again, this result stems from the persistence of the ECTs and the fact that it takes more than one period (month) to correct the temporary mispricing. This feature is in stark contrast with Dong et al. (2022), and it highlights the importance of modeling both prices and returns.

1.3 The Empirical Framework

This section describes the empirical setting, which is guided by the econometric framework of Section 1.2. I introduce a model to decompose the cross section of prices into a common permanent trend component and an idiosyncratic stationary one. Then, I describe the dimensionality reduction techniques I use to extract information from the cross section of error correction terms. I finally describe the out-of-sample procedure I follow.

1.3.1 Estimating the Trends

The first step in my empirical analysis is to decompose prices into a permanent and a transitory component, as in Section 1.2. The former is common across assets, while the latter is idiosyncratic. There are two potential approaches to estimate the common stochastic trends. The first one starts from a set of factors for the cross-section of returns, and it cumulates them to obtain the non-stationary trends (Favero et al., 2021). The second approach estimates the permanent components directly from the cross-section of prices (Banerjee and Marcellino, 2009). In this section I introduce both methods, and I explain why I will mostly rely on the latter for the empirical analysis.

Favero et al. (2021) introduce a way to estimate the common stochastic trends. Consider a candidate factor model for the cross-section of returns, such as the 5-factor model by Fama and French (2015), and denote the factors by f_t. Proxies for the common trends $\log F_t$ can be obtained by cumulating the factors, i.e. $\log F_t = \log F_{t-1} + f_t$. From an economic point of view, $\log F_t$ contains the values of buy-and-hold portfolios that replicate the factors f_t. Favero et al. (2021) show that these integrated factors are cointegrated with the prices of the long-short anomaly portfolios, giving rise to the error correction mechanism described in Section 1.2. This procedure is in line with my theoretical framework, but it has two main drawbacks.

First, this method does not allow the long and short legs of the anomaly portfolios to be driven by different trends. Indeed, the trends are built by cumulating factors which are designed to explain the cross-section of long-short portfolios, not of the long and short legs separately. As a consequence, integrating these factors may not be appropriate to capture the common drivers of my test assets. A possible solution could be to develop a set of factors for the long and short returns separately (e.g. with principal components), and to later cumulate them. However, this solution does not perform optimally, which brings me to the second drawback of this trend estimation method: factors built for returns are not guaranteed to describe prices. A key assumption of the empirical analysis is that the portfolio prices must be cointegrated with the trends, so that their mispricing shocks are stationary. This condition may not be satisfied by the cumulated factors, especially as the number of assets increases.

The approach I will use to estimate the common stochastic trends follows Bai (2004) and Banerjee and Marcellino (2009)'s Factor-augmented Error Correction Model (FECM). This method is useful to study large datasets with cointegration relationships which are unknown or difficult to model explicitly, as in this empirical setting. In particular, consider a cross-section of assets labeled by $i = 1, ..., N$ with I(1) log-prices $\log P_{i,t}$. Suppose the data generating process for the prices is

$$\log P_{i,t} = \lambda_i' \log F_t + u_{i,t}, \tag{1.19}$$

$$\log F_t = \log F_{t-1} + \nu_t, \tag{1.20}$$

with $u_{i,t}$ being I(0), $\log F_t = (\log F_{1,t}, ..., \log F_{r,t})'$, and ν_t is a vector of I(0) processes that drive the stochastic trends. Bai (2004) proves that the principal components of $\log P_{i,t}$ consistently estimate $\log F_t$ as N diverges[2]. Moreover, the number of factors r can be consistently estimated using information criteria (the Integrated Panel Criteria). I thus estimate the common trends $\log F_t$ as the cross-sectional principal components of the portfolio prices. I do so separately for the long and the short legs, ending up with the two set of factors $\log F_{L,t}$ and $\log F_{S,t}$ respectively.

Equation (1.19) describes a factor model for prices, rather than returns. Estimating trends directly from prices delivers a larger number of cointegrated portfolios. This feature is key for the empirical strategy, which relies on the stationarity of the mispricing shocks. The method also proves to be flexible enough to account for the different trends driving the long and short legs. Therefore, I will use it as the baseline approach to estimate the common stochastic trends.

[2]To be precise, the principal components consistently estimate the space spanned by $\log F_t$.

The literature on the predictability of returns has proposed other cointegrating frameworks. As an example, Campbell and Shiller (1988) use the assumption that prices and dividends are cointegrated in a dynamic dividend growth model. Lettau and Ludvigson (2001) model the cointegrating relation between consumption, income and wealth, with implications for the predictability of excess stock returns. Lettau and Ludvigson (2005), Bansal et al. (2009) and Bansal and Kiku (2011) model a relation between consumption and dividends, useful to model excess returns in the long-run. As I already mentioned, Favero et al. (2021) propose to cumulate the returns of cross-sectional factors to proxy for the common stochastic factors in a cross-section of prices. The papers mentioned above model relationships at the "aggregate" level (e.g. for the aggregate equity market). As I will show later in the empirical analysis, it is important to allow for different trends between the long and short components of anomaly portfolios, a feature which comes with the cointegrating framework I proposed above. Indeed, since the prices of long-minus-short portfolios are non-stationary, we need to allow for a difference in the permanent components of their legs (which would otherwise cancel out). The other cointegrating framework I mentioned, working at the aggregate level, would not allow for such flexibility. They could allow for different permanent components only through different loadings of the long and short legs on the common trends.

1.3.2 Predicting the Market

After estimating the common stochastic trends and thus the error correction terms for each asset, I use them to predict the market returns. In particular, I run regressions in the form of Equation (1.16) and (1.13), comparing the predictive performance of ECTs and lagged returns. The baseline regression I run is

$$r_{M,t,t+h} = \alpha + \delta f_{t,t+h} + \sum_{i=1}^{N} \beta_{i,L} u_{i,L,t} + \sum_{i=1}^{N} \beta_{i,S} u_{i,S,t} + \epsilon_{t+h}, \tag{1.21}$$

in which $r_{M,t,t+h}$ is the log-return of the market portfolio between time t and $t+h$; $f_{t,t+h} = \log F_{t+h} - \log F_t$ is the first difference of the common stochastic trends; $u_{i,L,t}$ ($u_{i,S,t}$) is the ECT of the long (short) leg of portfolio i at time t; N is the number of portfolios in the cross-section ($N = 100$ in the baseline regressions). In some specifications I will simply use the long or short legs, in which case I will use $\sum_{i=1}^{N} \beta_{i,L} u_{i,L,t}$ or $\sum_{i=1}^{N} \beta_{i,S} u_{i,S,t}$ as predictive terms, respectively.

The regression in Equation (1.21) does not only contain predictive terms, as $f_{t,t+h}$ is not observed at time t. I include this term in order to better identify the predictive coefficients $\beta_{i,L}$ and $\beta_{i,S}$, and to get closer to the econometric framework. Indeed, FECM models are particularly sensitive to omitted variables (Banerjee and Marcellino, 2009; Banerjee et al., 2014). In the out-of-sample exercises I will make sure to only use information up to time t, and I will substitute $f_{t,t+h}$ with its historical average to forecast the future value of the market return. This last step is consistent with the data generating process of Equation (1.20).

Since the number of portfolios N is large (100 in the baseline regressions), estimating Equation (1.21) with Ordinary Least Squares (OLS) would lead to overfitting and over-reacting to noise. Indeed, OLS maximizes the in-sample fit by construction, which can easily lead to poor out-of-sample performance. It is thus important to use methods to extract a signal in the high-dimensional setting and to protect against overfitting. I use two of the shrinkage techniques in Dong et al. (2022), principal components and Elastic Net.

Principal Component I consider the first principal component of the predictors $u_{i,l,t}$ with $l = L, S$, and I use it as a regressor alongside $f_{t,t+h}$ (which is left unchanged). Whenever the regressions are restricted to the sub-samples of long or short legs, I consider the first principal component as well. The advantage of this method is that the information is condensed in a single series. It is thus possible to analyze the recursive estimates of the slope coefficients and their evolution over time, which would be complicated in a ENet framework. This feature will be useful for the in-sample analysis. I will refer to this method as U-PCA.

ENet I estimate the predictive regression of Equation (1.21) via Elastic Net (Zou and Hastie, 2005). ENet combines both ℓ_1 and ℓ_2 penalties from lasso and ridge models. It still shrinks some coefficients to zero due to the ℓ_1 penalty, but it outperforms the lasso (e.g. when there are high correlations among predictors). I modify the standard ENet and I apply the shrinkage only to the predictors $u_{i,l,t}$ with $l = L, S$, leaving the regressors $f_{t,t+h}$ untouched. I follow Dong et al. (2022) and I use a modified Aikaike information criterion to select the value of the regularization parameter (Flynn et al., 2013). I will refer to this method as U-ENet.

1.3.3 Forecast Evaluation

I evaluate the forecasting performance of the ECTs using the the Campbell and Thompson (2008) out-of-sample R^2, henceforth R^2_{OS}. The R^2_{OS} statistics compares the forecast errors of a candidate model to a benchmark. In particular,

$$R^2_{OS} = 1 - \frac{MSFE_1}{MSFE_0}, \tag{1.22}$$

in which $MSFE_1$ ($MSFE_0$) is the Mean Squared Forecast Errors of the candidate (benchmark) model. $R^2_{OS} > 0$ implies that the candidate model has a superior forecasting performance than the benchmark. I follow the asset pricing literature and I use the the trailing mean of the market excess returns as benchmark model (Welch and Goyal, 2008). Given my forecasting horizon, which ranges from 1 to 12 months, this is the most appropriate benchmark to assess out-of-sample predictive performance (Goyal et al., 2023).

I test the null hypothesis $H_0 : R^2_{OS} \leq 0$ against the alternative $H_A : R^2_{OS} > 0$ using the Clark and West (2007) procedure. Let $\hat{r}^0_{t-1,t}$ and $\hat{r}^1_{t-1,t}$ be the forecasts of the market return under the benchmark and candidate model respectively. The forecast errors $\hat{e}_{0,t}$ and $\hat{e}_{1,t}$ can be obtained as

$$\hat{e}_{j,t} = r_{M,t-1,t} - \hat{r}^j_{M,t-1,t}, \quad j = 0, 1. \tag{1.23}$$

Clark and West (2007) propose to fit the regression

$$\hat{e}^2_{0,t} - \hat{e}^2_{1,t} + \left(\hat{r}^0_{M,t-1,t} - \hat{r}^1_{M,t-1,t}\right)^2 = \mu + \epsilon_t \tag{1.24}$$

via OLS and to do a t-test of $H_0 : \mu \leq 0$ against $H_A : \mu > 0$. Since $\mu \leq 0$ is equivalent to $MSFE_0 \leq MSFE_1$, this procedure let us test $H_0 : R^2_{OS} \leq 0$. I compute t-statistics using heteroskedasticity and autocorrelation consistent (HAC) standard errors (Newey and West, 1986).

As a final remark, I will inspect not only the statistical significance of the R^2_{OS}, but also its magnitude. Campbell and Thompson (2008) show that a monthly R^2_{OS} of 0.5% is already economically significant. Therefore, I will always compare my estimates with this threshold.

1.3.4 Economic Value

As outlined in Section 1.3.3, I mostly evaluate competing models according to their R^2_{OS}. I complement this approach by analyzing their economic value to investors. I perform a portfolio allocation exercise and I compute the utility gains for a risk averse agent who can invest in the market portfolio and a risk-free asset.

Consider a risk-averse investor who maximizes

$$\max_{w^j_{t,t-1}} w^j_{t,t-1}\hat{r}^j_{M,t,t-1} - \frac{1}{2}\gamma\left(w^j_{t,t-1}\right)^2 \hat{\sigma}^2_{M,t,t-1}, \quad j = 0, 1, \tag{1.25}$$

where γ is the parameter governing the degree of relative risk aversion; $w^j_{t,t-1}$ is the weight given to the market portfolio according to forecasting model j, with $j = 0$ ($j = 1$) referring to the benchmark (competing) model[3]; $\hat{\sigma}^2_{M,t,t-1}$ is the forecast of the variance of the market excess return, that I assume to be the same across the two forecasting strategies.

The optimal portfolio weight under model j is

$$w^{j,*}_{t,t-1} = \frac{1}{\gamma} \frac{\hat{r}^j_{M,t,t-1}}{\hat{\sigma}^2_{M,t,t-1}}. \tag{1.26}$$

Given $w^{j,*}_{t,t-1}$, I compute the returns realized by the investor's portfolio in an out-of-sample exercise. I denote the mean and variance of such realized returns as $\bar{r}_j$ and $\hat{\sigma}^2_j$ respectively. The average utility realized by the investor is

$$\bar{U}_j = \bar{r}_j - \frac{1}{2}\gamma\hat{\sigma}^2_j, \quad j = 0, 1. \tag{1.27}$$

The economic value of the competing forecasting strategy over the benchamrk is estimated via the average utility gain

$$\Delta U = \bar{U}_1 - \bar{U}_0. \tag{1.28}$$

ΔU can be thought of as the fee that the investor would be willing to pay in order to access the forecasts of the competing model.

I follow the literature and I use a "mixed" scheme to perform the asset allocation exercise (McCracken and Valente, 2018). That is, the investor uses a recursive window to compute $\hat{r}^j_{M,t,t-1}$, but estimates $\hat{\sigma}^2_{M,t,t-1}$ as the sample variance of the market excess return over a 60-months rolling window. I consider a risk aversion coefficient equal to 3. I rule out extreme portfolio weights by constraining them between -1 and 2.

[3]While the investor allocates $1 - w^j_{t,t-1}$ to the risk-free asset.

1.4 Data

In the empirical analysis I use the cross-section of long-short anomaly portfolios of Dong et al. (2022). This dataset includes 100 portfolios that are representative of anomalies from the cross-sectional asset pricing literature. For each anomaly, these long-short portfolios are built by sorting stocks into decile portfolios based on a given characteristic. The anomaly portfolio goes long (short) the tenth (first) decile portfolio. As an example, if we sorted stocks according to their book-to-market ratio, the anomaly portfolio would go long the value stocks, and short the growth ones, reproducing the value-growth anomaly. I will consider separately the long and the short legs of these portfolios, so that in total I have a cross-section of 200 assets. In the case of the book-to-market ratio, I end up considering the value and the growth stocks as two separate portfolios.

The sample period goes from 1970:01 to 2017:12 at the monthly frequency. I present some summary statistics for the returns in Table I. (It mirrors Table I in Dong et al. (2022), I report it here for convenience). We can see that most of the portfolios are indeed anomalies, i.e. they generate statistically significant alphas with respect to the Fama and French (1993) three factor model. As an example, 75 anomalies out of 100 have an alpha which is positive at the 10% confidence level.

The correlation across the returns of the long (short) legs of the anomalies is on average 0.76 (0.86), a large value. On the contrary, the correlation for the long-short returns is low on average (0.08). Dong et al. (2022) interpret these facts as consistent with a common component in the returns of the long and short legs, which cancels out in the long-short portfolios. I argue that these statistics are misleading, as they only focus on returns. The empirical analysis of Section 3.4 will show that a different picture emerges when considering prices. I plot the (log)prices of the long, short and long-short portfolios in Figure 1.1. Euristically, the long and short legs seem to be driven by different trends, and the long-short anomaly portfolio prices are not stationary.

Table I

Summary Statistics

The table reports summary statistics for monthly anomaly portfolio returns for 100 anomalies. The sample period is 1970:01 to 2017:12. For each anomaly, stocks are sorted into value-weighted decile portfolios according to the characteristic underlying the anomaly. The long-short anomaly portfolio goes long (short) the tenth (first) decile portfolio.

Number of anomalies	100
Fama and French (1993) three-factor model alpha	
Number of long-short anomaly portfolio returns with $\lvert t\text{-stat.}\rvert \geq 1.645$	75
Number of long-short anomaly portfolio returns with $\lvert t\text{-stat.}\rvert \geq 1.96$	71
Number of long-short anomaly portfolio returns with $\lvert t\text{-stat.}\rvert \geq 2.58$	56
Number of long-short anomaly portfolio returns with $\lvert t\text{-stat.}\rvert \geq 3$	49
Average correlation across anomaly decile rankings 0.05	
Average correlation across monthly anomaly excess returns	
Long leg	0.76
Short leg	0.82
Long-short	0.08
Long-leg anomaly portfolio excess returns	
Average of sample means	0.71%
Average of sample standard deviations	5.16%
Short-leg anomaly portfolio excess returns	
Average of sample means	0.33%
Average of sample standard deviations	6.20%
Long-short anomaly portfolio returns	
Average of sample means	0.38%
Average of sample standard deviations	4.37%

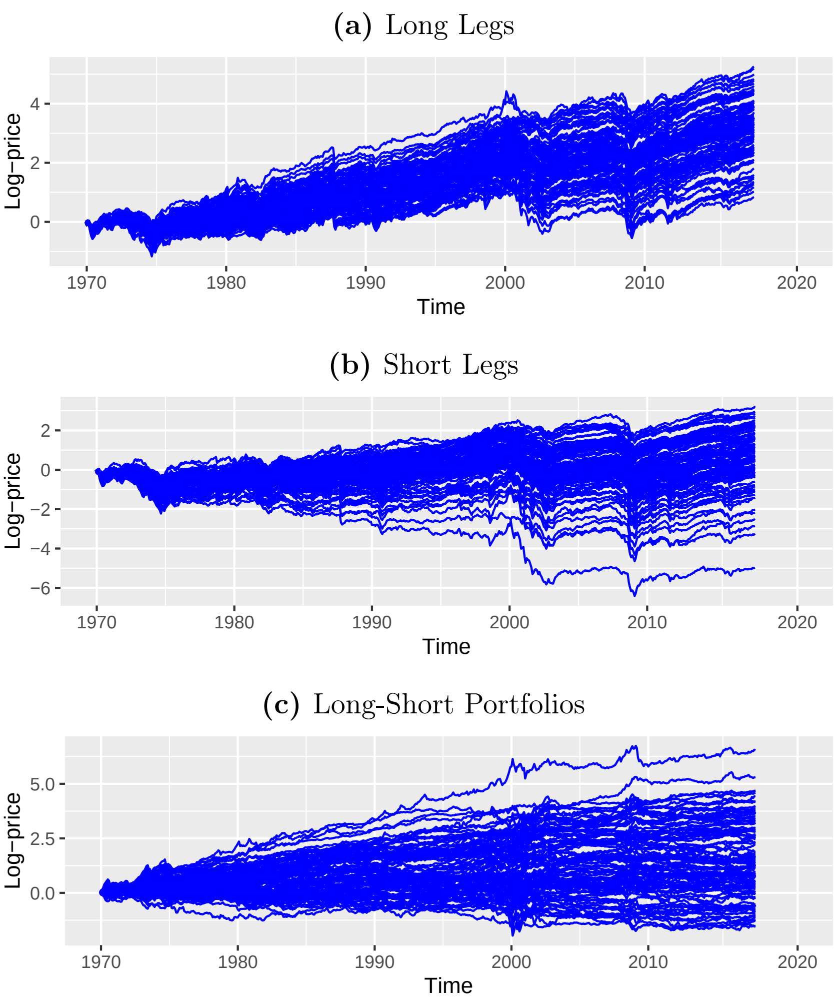


Figure 1.1: Each panel depicts the log-prices of the (a) long, (b) short, (c) long-short anomaly portfolios.

1.5 Empirical Results

This section presents the empirical results. I start by describing the in-sample analysis. I later turn to the out-of-sample exercise, which is the bulk of my findings. Finally, I show the results at multiple forecasting horizons.

1.5.1 In-Sample

I start by describing the results using the full sample, from 1970:01 to 2017:12. This section focuses mainly on the estimated common trends and the properties of the error correction terms. This cointegration analysis is important to see whether the error correction mechanism of Section 1.2 is present in the data. I will also provide evidence for asymmetric limits of arbitrage, focusing on the persistence of the long and short portfolio ECTs. Finally, I discuss the estimates of the predictive coefficients in the principal components framework.

I estimate the common stochastic trends of (1.19) using the cross-sectional principal components of the log-prices. I use information criteria to obtain the number of factors. I apply the information criteria in Bai (2004) to the data in levels, and the ones in Bai and Ng (2002) to the differenced series. This procedure captures both the I(1) and I(0) factors, which is important as omitting relevant variables in FECM models is problematic (Banerjee and Marcellino, 2009). The long and short legs are driven by 3 factors each, while the cross-section of both long and short legs' prices (L&S for brevity) requires 5 factors.

The larger L&S cross-section requires more drivers. This fact is consistent with the long and short legs being driven by different trends. Were this not the case, we would expect to find only 3 factors when considering them together. This simple finding also sheds new light on the evidence of Favero et al. (2021). The authors show that the prices of the long-short anomaly portfolios are not stationary, which is consistent with *i*) heterogeneous trends in the long and short legs *ii*) heterogeneous exposures of the legs to the same trends. Having more factors in the L&S cross-section brings new evidence in favor of the former hypothesis.

I plot the common trends for the long, short, long and short portfolios in Figure 1.2. All the drivers appear to be quite persistent, as they are constructed from the cross-section of prices. Once again, we can visually see that the factors that drive the two legs are different. By inspecting panel (c), which includes both types of portfolios, it appears that the two only share one common factor, the one depicted in red.[4] This is to be expected,

[4]This common trend is basically the cumulative aggregate market, as the two series share a 99% correlation.

as the information criteria estimated 5 rather than 3 + 3 factors. After estimating the

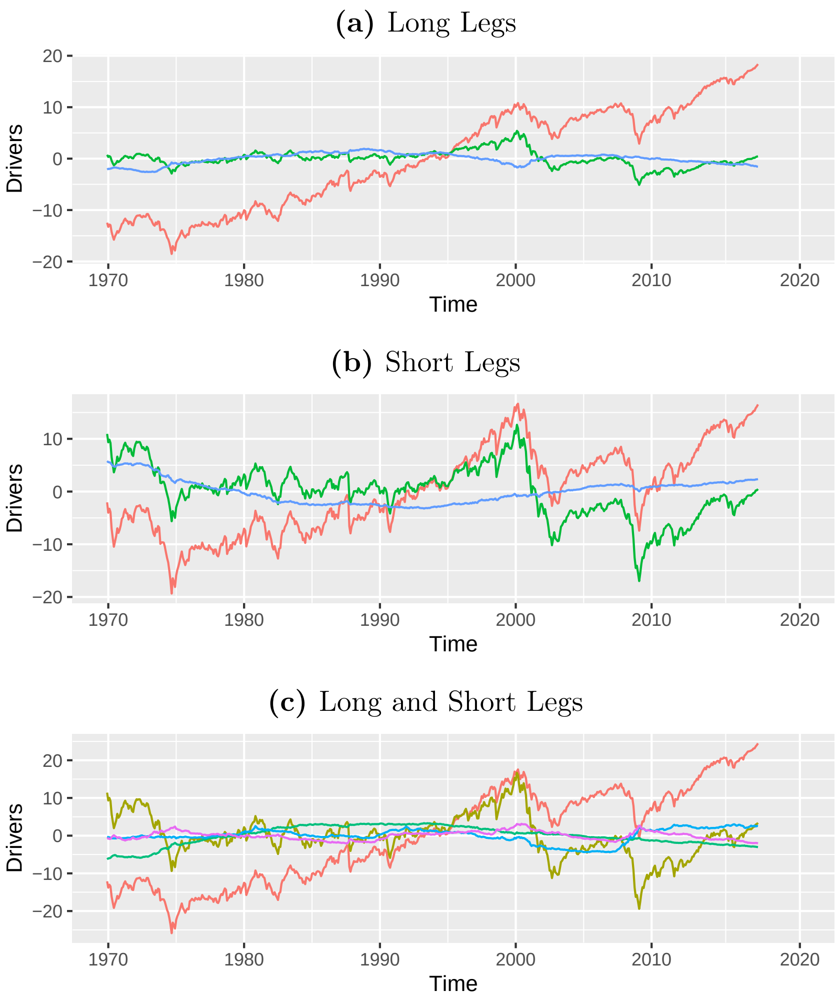


Figure 1.2: Each panel depicts the common cross-sectional trends for the log-prices of the (a) long, (b) short, (c) long and short legs.

trends, I compute the error correction terms $\hat{u}_{i,L,t}$ and $\hat{u}_{i,S,t}$ by fitting Equation (1.19) for each portfolio leg. In line with the framework of Section 1.2, I estimate the persistence of each ECT series, obtained as the autoregressive coefficient of an AR(1) model. I show their histogram in Figure 1.3. Interestingly, the persistence distribution is different for the long and short portfolios: the latter have more counts in the bins close to one. This fact is consistent with the asymmetric limits of arbitrage of Assumption 1. Moreover, I run a statistical test on the means of the two distributions. The mean persistence of the long legs is smaller than for the short legs, at the 10% level. This fact is robust to changing the number of drivers used to compute the mispricings.

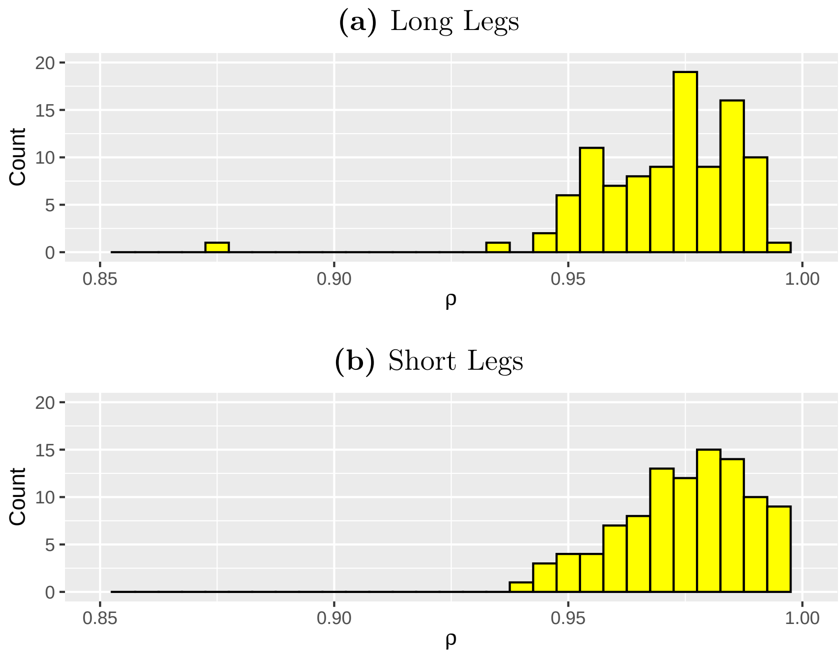


Figure 1.3: The figure shows histograms for the persistence ρ of the error correction terms $u_{i,L,t}$ and $u_{i,S,t}$ for the (a) long and (b) short legs respectively.

Finally, I look at the recursive estimates of the slope coefficients in the predictive regressions, i.e. $\beta_{i,L}$ and $\beta_{i,S}$ in Equation (1.21). I use the dimension reduction methods of Section 1.3.2 to avoid overfitting, since the number of regressors is large. The principal component method is more suitable than ENet for an in-sample analysis. Indeed, suppose to fit the regression

$$r_{M,t,t+h} = \alpha_l + \beta_l u_{l,t}^{PCA} + \epsilon_{l,t+h}, \tag{1.29}$$

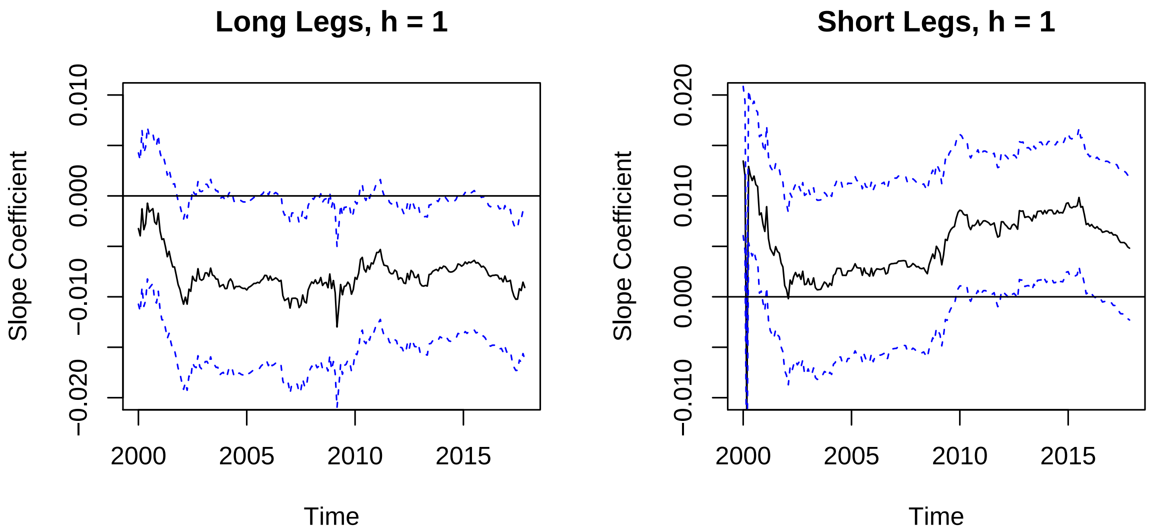
with $u_{l,t}^{PCA}$ being the first principal component of the long ($l = L$) or short ($l = S$) portfolios ECTs. This approach delivers a single slope coefficient to compare across the long and short sample, i.e. β_L and β_S. On the contrary, ENet shrinks the slope coefficients to zero, but in general more than one coefficient will survive. It is difficult to interpret the results if the selected coefficients for the long and short samples belong to different portfolios. I will thus focus on the principal component method.

Figure 1.4 shows the recursive estimates of the slope coefficient used to compute one-month ahead forecasts ($h = 1$). It includes 90% confidence intervals, computed according to Bai and Ng (2006). In the case of the long (short) legs, I set the sign of the principal component so that it is positively correlated with $1/N \sum_{i=1}^{N} u_{i,L,t}$ $(1/N \sum_{i=1}^{N} u_{i,S,t})$. In this way, an increase in the principal component can be interpreted as a general increase in the ECTs.

The slope coefficients of the long and short legs have opposite signs. The estimates for the long legs are negative and statistically significant towards the end of the sample. They are also fairly stable over the estimation window. On the contrary, the slopes of the short legs are mostly insignificant and, if anything, positive. They also show a shift around 2010, which may signal a structural break due to the Financial crisis. Consistently with Proposition 1, the slope coefficients for the long leg β_L are negative, statistically significant, and larger (in absolute value) than β_S. These findings are in line with the larger persistence of the short portfolio ECTs, as documented in Figure 1.3. This latter fact may also explain the positive estimates of β_S after 2010, which are at odds with the error correction framework of Section 1.3.2: if the short portfolio prices were cointegrated, we would expect their deviations from equilibrium to negatively predict the market.

1.5.2 Out-of-Sample

This section describes the results of the out-of-sample analysis, which is the most appropriate way to test market efficiency questions (Welch and Goyal, 2008; Martin and Nagel, 2021). I use the period 1970:01 to 1999:12 as the initial in-sample estimation period. I later use expanding estimation windows and the out-of-sample period is thus 2000:01 to 2017:12. I present results at multiple forecasting horizons, highlighting the advantages of focusing on prices rather than returns. Throughout my out-of-sample analysis I will use the historical mean as benchmark, which is the relevant benchmark given my forecasting horizon. Indeed, recent empirical evidence shows that it is hard to find predictors that



(a) Long legs **(b)** Short legs

Figure 1.4: The solid line depicts the recursive estimates of the slope coefficients of the principal component forecasts (one-month ahead) from the (a) long and (b) short ECTs. The dashed lines denote 90% confidence intervals.

consistently outperform the mean (Goyal et al., 2023).

I start with one-month ahead forecasts ($h = 1$) of the market excess returns. The forecasts are based on the estimates of Equation (1.21) using either the first principal components (U-PCA) or ENet (U-ENet) on the error correction terms. I conduct the out-of-sample exercise on three separate sample, first considering all the portfolios, and then focusing on the long and short ones separately. I compare the results with forecasts based on simple lagged returns as in Dong et al. (2022). For the sake of comparability, I use also in this case the first principal component (R-PCA) or ENet (R-ENet) on the returns for the three portfolio samples. I will also conduct the same out-of-sample exercise using the cross-section of long-short anomaly portfolios. As outlined in Section 1.3.3, I evaluate the forecasting performance of the different approaches via the Campbell and Thompson (2008) out-of-sample R^2 (R^2_{OS}).

I present the results in Table II. In the first row I use as regressors the error correction terms of both the long and the short legs (L&S). Columns (1) and (2) are obtained using

Table II

R^2_{OS} Statistics

The table reports Campbell and Thompson (2008) out-of-sample R^2 (R^2_{OS}) in percent for market excess return forecasts one month in the future. The out-of-sample period is 2000:01 to 2017:12. U-PCA (R-PCA) forecasts are based on the first principal component of the cointegrated error correction terms (returns) of the portfolios. U-ENet (R-ENet) forecasts are built from elastic net predictive regressions that include all the cointegrated error correction terms (returns) of the portfolios. L&S, L, S, LMS refer to the cross-section of long and short, long, short, long minus short portfolios. *, ** and *** indicate significance at the 10%, 5% and 1% level respectively for the positive R^2_{OS}, based on the Clark and West (2007) test.

	(1) U-PCA	(2) U-ENet	(3) R-PCA	(4) R-ENet
L&S	0.56**	-1.88	1.20	-0.88
	(1.86)	(-0.35)	(1.19)	(1.53)
L	0.74*	1.41**	0.79	-1.85
	(1.63)	(1.68)	(0.98)	(0.96)
S	-0.63	-2.08	1.57*	4.00*
	(-1.24)	(-0.39)	(1.34)	(1.37)
LMS	-1.01	-1.15	1.33*	2.46**
	(-2.10)	(0.34)	(1.35)	(1.86)

the U-PCA and U-ENet methods respectively. The former approach delivers an R^2_{OS} of 0.56%, which is both statistically (at the 5% level) and economically significant. Indeed, it is larger than the 0.5% threshold of Campbell and Thompson (2008) for economic significance. The U-ENet method, instead, does not beat the historical mean benchmark ($R^2_{OS} = -1.88\%$). This fact is probably due to an overfitting of the signal from the short portfolio ECTs, as will be clearer later on. Columns (3) and (4) refer to the R-PCA and R-ENet approaches respectively. Using both the long and short leg returns does not improve the forecasting performance with respect to the historical mean. R-PCA delivers a large but insignificant R^2_{OS} of 1.20%, which is negative under the R-ENet specification (-0.88%). The lack of statistical significance for the long and short returns is probably due to overfitting the former, as I will discuss shortly.

I run the forecasting exercise on the subsamples of the long and short portfolio legs, to disentangle their respective contributions. The results can be found in the second (L) and third (R) rows of Table II respectively. In the case of the long portfolios, both the U-PCA and U-ENet methods deliver economically and statistically significant out-of-sample R^2. U-ENet outperforms U-PCA, with an R^2_{OS} of 1.41% against 0.74%. In line with Proposition 2, constructing the forecasts from the long leg returns delivers weaker results. The R-PCA approach gives an out-of-sample R^2 of 0.79%, which is large but not statistically significant. R-ENet is instead beaten by the historical mean (-1.85%). The deviations of prices from equilibrium are key to predict the market, and the information they carry is lost when only the returns are considered.

The picture is reversed for the short legs. The third row of Table II shows that neither U-PCA nor U-ENet outperforms the historical mean. They both have negative R^2_{OS} , -0.63% and -2.08% respectively. This finding is consistent with Proposition 1. The short leg ECTs are more persistent than the long ones, due to asymmetric limits of arbitrage, and they are worse predictors. Interestingly, the short returns are useful to predict the market. The R-PCA and R-ENet methods achieve an R^2_{OS} of 1.57% and 4%.[5]

The fourth row of Table II presents the results on the cross-section of long-short anomaly portfolios. From the econometric framework, there is no reason to expect that the ECTs of these portfolios lead to market predictability. Indeed, the i-th anomaly portfolio deviations from equilibrium are in the form $u_{i,L,t} - u_{i,S,t}$. The latter expression is a noisy signal for $u_{i,L,t}$ or $u_{i,S,t}$ which enter the expression of the market return (see e.g. Equation (1.7)). U-PCA and U-ENet consistently deliver a negative R^2_{OS}, -1.01% and -1.15% respectively. Instead, the long-short anomaly returns predict the market as in Dong et al. (2022).[6]

[5]The latter number is considerably larger than the one in Dong et al. (2022), due to different out-of-sample periods. Indeed, the authors start their out-of-sample exercise in 1985:01. I consider a shorter window because my analysis relies on long-run relations, that need more observations to be estimated.

[6]Computing the prices of the long-minus-short anomaly portfolios either as the prices of long legs minus the prices of short legs, or by cumulating the long-minus-short portfolio returns, virtually does not

1.5.3 Multiple Horizons

So far I have described the results for the one-month ahead forecasts. This section presents the analysis for horizons of one quarter and one year in the future. According to the econometric framework of Section 1.2, I expect the error correction terms to provide valuable information at multiple forecasting horizons.

I present the forecasting results at multiple horizons in Table III. The first row replicates Table II. For the sake of brevity, I do not report the forecasts based on the portfolio returns, whose R^2_{OS} are small and insignificant beyond the one month horizon (Dong et al., 2022). If we consider both the long and the short legs (L&S), we can see that the ECTs contain information to predict the market at multiple horizons. Under the U-PCA method, the R^2_{OS} at the 3 and 12 months horizon is 1.21% and 1.36% respectively. Both figures are statistically and economically significant. Even though the U-ENet forecasts are beaten by the historical average at the quarterly horizon, they deliver a remarkable 14.87% one year in the future.

If we focus on the long (L) and short (S) leg subsamples, a similar picture emerges. In the case of the long legs, both U-PCA and U-ENet deliver positive and large R^2_{OS} at the quarterly and annual horizon. The statistics get larger as the horizon increases, in line with Proposition 3: since the mispricings are not corrected in just one period, they contain valuable information to forecast at longer horizons. On the contrary, the short legs do not achieve predictability at any horizons.

In light of these findings, I compare the recursive estimates of the slope coefficients for the long and short ECTs for one-year ahead predictive regressions. The case of one-quarter ahead forecasts is similar and is not reported for the sake of brevity. As already explained in Section 1.5.1, I only focus on the principal component regressions to interpret the results. Figure 1.5 depicts the slope estimates, together with 10% confidence intervals. In line with Proposition 3, the long legs coefficients β_L for the annual horizon forecasts

affect the results.

Table III

R^2_{OS} Statistics - Multiple Horizons

The table reports Campbell and Thompson (2008) out-of-sample R^2 (R^2_{OS}) in percent for market excess return forecasts at horizons of one month, one quarter and one year. The out-of-sample period is 2000:01 to 2017:12. U-PCA forecasts are based on the first principal component of the cointegrated error correction terms of the portfolios. U-ENet forecasts are built from elastic net predictive regressions that include all the cointegrated error correction terms (returns) of the portfolios. L&S, L, S refer to the cross-section of long and short, long, short portfolios. *, ** and *** indicate significance at the 10%, 5% and 1% level respectively for the positive R^2_{OS}, based on the Clark and West (2007) test.

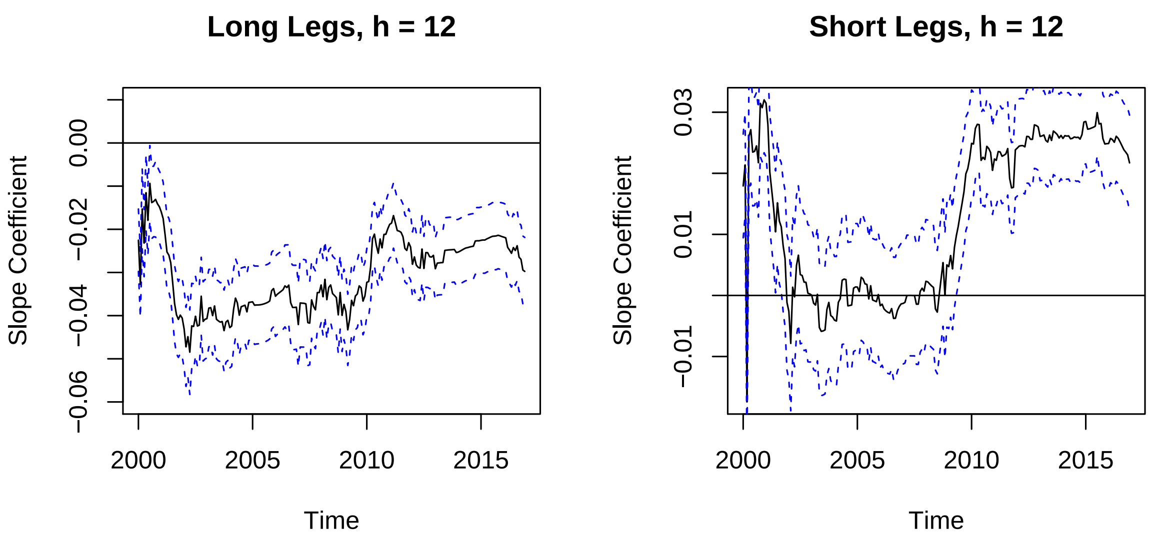
	L		S		L&S	
	U-PCA	U-ENet	U-PCA	U-ENet	U-PCA	U-ENet
$R_{M,t,t+1}$	0.74*	1.41**	-0.63	-2.08	0.56**	-1.88
	(1.63)	(1.68)	(-1.24)	(-0.39)	(1.86)	(-0.35)
$R_{M,t,t+3}$	1.38**	2.07**	-2.38	-8.06	1.21***	-7.54
	(1.87)	(1.76)	(-1.86)	(-1.26)	(2.33)	(-0.79)
$R_{M,t,t+12}$	2.81***	13.10***	-10.12	-11.49	1.36**	14.87***
	(2.28)	(4.29)	(-3.34)	(-1.46)	(1.67)	(5.04)

are larger (in magnitude) than for the monthly case of Figure 1.4. The estimates are always negative and statistically different from zero. Concerning the short legs, until 2010 their slope coefficients are non-significant and clearly smaller than in the long leg case. However, after 2010 there appears to be a structural break, that leads to a shift of the estimated coefficients. Once again, this fact may hint at a structural break due to the Financial crisis, which would deserver further investigation.

1.5.4 Economic Value

In this section, I compare the forecasting models in terms of their economic value to a risk-averse investor. This procedure tests the robustness of the out-of-sample results. An asset allocation exercise also provides a direct measure of the economic significance of the results, which was so far assessed via the threshold of $R^2_{OS} \geq 0.5\%$.

I compute the utility gains for an investor with a risk aversion coefficient of 3 following



(a) Long legs **(b)** Short legs

Figure 1.5: The solid line depicts the recursive estimates of the slope coefficients of the principal component forecasts (one-year ahead) from the (a) long and (b) short ECTs. The dashed lines denote 90% confidence intervals.

the procedure in Section 1.3.4. I perform the asset allocation exercise at horizons of one month, one quarter and one year in line with the previous sections. I will compare the portfolios built starting from the ECT-based forecasts (U-PCA and U-Enet) with the historical mean. The utility gains are always expressed as an annualized percentage. They can be interpreted as the management fee an investor would pay in order to construct the portfolio according to the ECT-based forecasts.

Table IV presents the results, overall corroborating the findings from the previous sections. If we consider only the long legs (L), both the U-PCA and U-ENet methods lead to sizable utility gains, at every investment horizon. The gains range from 1.22% to 6.38%. For each horizon, the U-ENet method systematically outperforms U-PCA, in line with the larger R^2_{OS} of Table III. If instead we consider the short legs (S), the utility gains are always negative or small, ranging from -2.72% to 0.14%. Using this subsample to forecast the market excess return, a mean-variance investor would have a larger utility by simply using the historical mean, in line with the previous findings.

Table IV
Utility Gains

The table reports the annualized average utility gains (ΔU) in percent for a risk-averse investor at horizons of one month, one quarter and one year. The out-of-sample period is 2000:01 to 2017:12. U-PCA forecasts are based on the first principal component of the cointegrated error correction terms of the portfolios. U-ENet forecasts are built from elastic net predictive regressions that include all the cointegrated error correction terms (returns) of the portfolios. L&S, L, S refer to the cross-section of long and short, long, short portfolios. The results are obtained by using a risk-averse coefficient of 3 and a 60-month rolling window to estimate the sample volatility.

	L		S		L&S	
	U-PCA	U-ENet	U-PCA	U-ENet	U-PCA	U-ENet
ΔU_1	1.77	5.06	−1.93	0.14	0.82	−0.61
ΔU_3	2.28	6.15	−2.72	−1.8	1.22	−0.73
ΔU_{12}	1.22	6.32	−1.93	−1.19	0.06	9.48

The results are mixed when considering the long and short legs together (L&S). The U-PCA method delivers utility gains that go from 0.06% to 1.22%, always lower than in the long legs case, but still positive. However, the U-ENet presents gains which are negative at the horizon of one and three months (−0.61% and −0.73% respectively), but largely positive at the annual horizon (9.48%). As already highlighted in the previous sections, this fact may be due to U-ENet overreacting to noise from the short legs, which are less likely to be stationary.

1.5.5 Additional Results and Robustness

Finally, I show additional results that qualify the behavior of the expected returns generated by the model. Following Bianchi et al. (2023), I impose an economic restrictions on the forecasts coming from the model. In particular, I impose that investors require a non-negative equity premium. That is, I impose that the predicted excess return is non-zero. I present the results in Appendix C, Table V. Most of the results coming from the U-PCA model are virtually unaffected, both in terms of magnitude of the R^2_{OS} and its statistical significance. This could be a sign that the predictions generated by the model are in line with economic theory. However, the predictions generated by the U-ENet model are worsened by such non-negativity constraints, even though mostly for the non-long-only cases.

I also test the robustness of my results by applying the same procedure to another dataset. In particular, I use the anomaly set from Jensen et al. (2023), who provide 154 portfolios, and apply the same statistical procedure described above. I present the results in Appendix C, Table VI. Due to the limitation of the dataset, I can only perform the analysis on the bare anomaly portfolios, and cannot divide them into their long and short components. However, extracting a signal still delivers smaller forecast errors than for the historical mean case, as show by all the R^2_{OS} measures for the U-PCA model. In the U-ENet case the results are less clean, even though they become stronger by increasing the forecasting horizon.

1.6 The Model

This section provides a microfoundation for the data generating process of Section 1.2. I present a standard model with slow-moving capital (Duffie, 2010), in which a mass of inattentive investors can only trade once every k periods. I show that such a model can account for temporary deviations of prices from the fundamental value, which are slowly corrected towards the equilibrium. I later generalize the setting to study the implications of asymmetric limits of arbitrage for price dynamics. I do so by introducing short selling costs, in the spirit of Gromb and Vayanos (2010). Departing from the previous literature, I specify the constraints in logistic form as to obtain analytical expressions. I show that the financial friction significantly affects the persistence of price deviations, as posited in the econometric framework.

1.6.1 Price Dynamics with Slow-Moving Capital

I start by describing the model of Duffie (2010), which I use to study the dynamics of price deviations from equilibrium. The model features both attentive and inattentive investors: the former can always buy and sell assets, while the latter only trade once every k periods. The inattentive agents do not react immediately to changes in the asset fundamentals, such as a permanent increase in its supply, thus causing the price to overshoot and depart from the fundamental value. However, these deviations are only temporary: as investors slowly change their positions in the asset to account for the new information, the price moves towards the new equilibrium value.

Consider an economy with two assets. Time is discrete and infinite. At time t there are Z_t units of a risky asset, whose price is P_t and which pays X_t dividends. There is a risk-free asset, whose gross return R is assumed to be constant and exogenous. The risky asset supply and dividends are exogenous and follow the vector autoregressive process

$$\begin{pmatrix} Z_t \\ X_t \end{pmatrix} = \Lambda \begin{pmatrix} Z_{t-1} \\ X_{t-1} \end{pmatrix} + \Sigma^{1/2}\epsilon_t, \tag{1.30}$$

in which Λ and $\Sigma^{1/2}$ are 2×2 matrices and $\epsilon_t = (\epsilon_{z,t}, \epsilon_{x,t})'$. $\Sigma^{1/2}$ is positive semi-definite and symmetric, while I impose no restrictions on Λ.

There are two types of agents with constant absolute risk aversion (CARA) utility. A mass $1-q$ of attentive investors trade assets in each period. Their aggregate demand for the risky asset at time t is K_t. Appendix B shows that their demand is given by

$$K_t = (1-q)\,\frac{\mathbb{E}_t\left[P_{t+1} + X_{t+1}\right] - RP_t}{\phi \mathbb{V}_t\left[P_{t+1} + X_{t+1}\right]}, \tag{1.31}$$

with ϕ being the harmonic mean of the attentive investors' risk-aversion coefficients. Equation (1.31) is the usual demand for risky assets under CARA utility.

There is a mass q of inattentive investors who can only trade once every k periods.

They cannot change their portfolio weights for k periods, and all the dividends from the risky asset are reinvested at the risk-free rate. Their activities are uniformly distributed across time, so that each period a fraction q/k can trade. I denote the inattentive investors' aggregate demand at time t with D_t. Appendix B derives

$$D_t = \frac{q}{k} \frac{\mathbb{E}_t \left[P_{t+k} + \sum_{n=1}^{k} R^{k-n} X_{t+n} \right] - R^k P_t}{\theta \mathbb{V}_t \left[P_{t+k} + \sum_{n=1}^{k} R^{k-n} X_{t+n} \right]}, \tag{1.32}$$

in which θ is the harmonic mean of the inattentive investors' risk-aversion coefficients.

At time t the total available supply of the risky asset is different from Z_t, as a mass of $q(1-k)$ inattentive investors cannot trade and their positions are held off the market. The market clearing condition therefore reads

$$K_t + D_t = Z_t - D_{t-1} - ... - D_{t-k+1}. \tag{1.33}$$

I collect the inattentive investors' positions currently unavailable for trade in the vector $H_t = (D_{t-1}, ..., D_{t-k+1})'$. Since H_t affects the net supply of the risky asset, the relevant state vector to find the equilibrium is $Y_t = (Z_t, X_t, H_t)'$. It can be proved that in such setting the risky asset price and demand are linear in the state vector, that is

$$P_t = c' Y_t, \tag{1.34}$$

in which c must be computed numerically.

I consider a simple case to show the intuition behind the model. I assume that at time $t = 1$ there is a shock which brings the supply from Z_0 to Z_1, with a conditional mean and variance of 0 and 0.1 respectively. The asset supply is constant after this shock, i.e. $Z_t = Z_1$ for $t \geq 1$. The dividends X_t are i.i.d., with 0 mean and variance equal to 0.1.

Referring to Equation (1.30), for $t \geq 1$ we have

$$\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \Sigma^{1/2} = \begin{pmatrix} 0 & 0 \\ 0 & \sqrt{0.1} \end{pmatrix}. \tag{1.35}$$

The fraction of inattentive investors is $q = 0.8$, their period of inactivity is $k = 32$ and they start with no asset holdings, that is $H_0 = 0$. The gross risk-free rate is $R = 1.01$ and the risk-aversion coefficients are $\phi = \theta = 1$. I consider a shock that brings the supply to $Z_1 = 2$.[7]

In this setting it is possible to compute analytically the "fundamental value" of the risky asset, V_t, which I define as the price at the steady state. In the spirit of Anufriev and Tuinstra (2013), I will focus on deviations of prices from this equilibrium value. Appendix B shows that

$$V_t = \frac{Z_t}{\frac{(1-q)(1-R)}{\phi\sigma_1^2} + \frac{q\left(1-R^k\right)}{\theta\sigma_k^2}}, \quad t \geq 1, \tag{1.36}$$

with the expressions of σ_1^2 and σ_k^2 provided in the Appendix.

Figure 1.6 shows the dynamics of price deviations from the equilibrium, that is $P_t - V_t$. I interpret this difference as a proxy for the error correction term of the empirical framework. After the massive shock at time $t = 1$, the supply of the risky asset increases dramatically. However, only 22.50% of investors are available to absorb this shock[8], implying a sizable price decline. Even though the price is below the fundamental value, in the following periods only a small fraction of investors can adjust their portfolios and buy it. One period at a time, inattentive demand increases, and the price fluctuations following the initial shock are dampened and fade away.

[7] Considering a time 0 conditional variance of 0.1, the shock is sizable.
[8] The number is obtained as $1 - q + q/k = 0.225$.

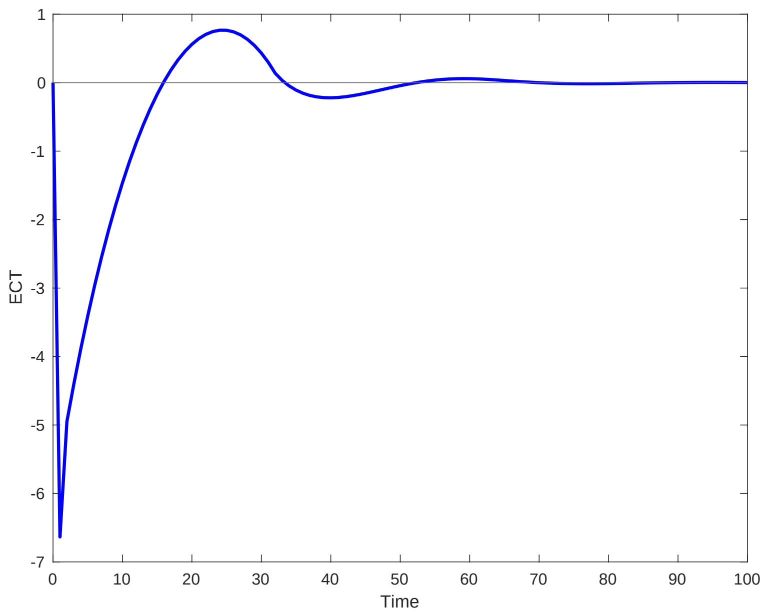


Figure 1.6: The figure depicts the price deviations from the fundamental value after a positive supply shock at time $t = 1$.

1.6.2 Short Selling Costs

This section builds on the framework of Duffie (2010) and adds a financial friction, to study its impact on the price dynamics. This latter feature brings the model closer to the empirical framework, in which I argued that asymmetric limits of arbitrage affect the persistence of mispricings.

I model asymmetric limits of arbitrage through short selling costs, in the spirit of Gromb and Vayanos (2010). If an agent holds x units of the risky asset, the holding costs are given by a function $\kappa(x)$. A natural choice would be to use $\kappa(x) = -\kappa|x|1_{\{x<0\}}$, in which $\kappa > 0$ is a constant and $1_{\{x<0\}}$ is the indicator function. Since this functional form complicates the analysis, I replace it with a logistic cost function

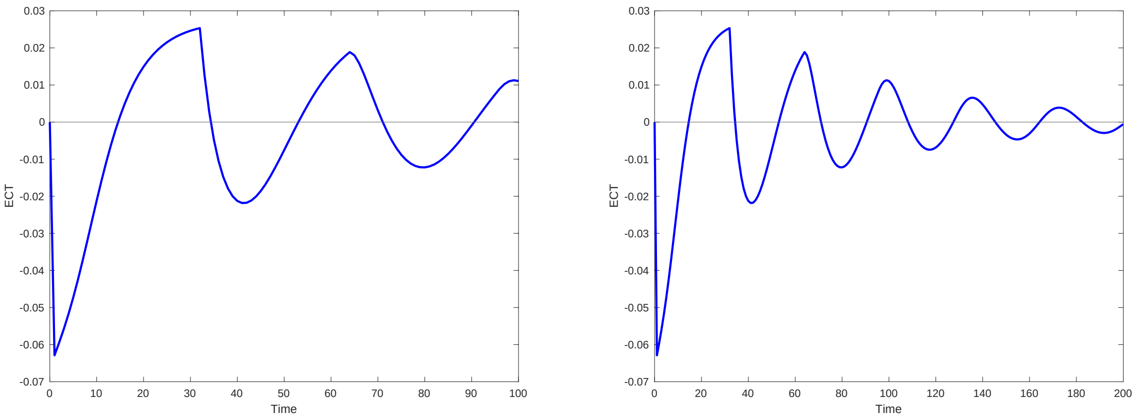
$$\kappa(x) = \kappa x \frac{e^{-\gamma x}}{1 + e^{-\gamma x}}. \tag{1.37}$$

The logistic costs in Equation (1.37) are positive for $x > 0$, which is a drawback. However, with a large enough value of γ, they provide an approximation of the usual step-function short selling costs. Figure 9 in Appendix B shows a comparison between the two functional forms.

I repeat the analysis of Section 1.6.1, adding short selling constraints in the form of the logistic costs of Equation (1.37). I keep all the parameters as before, and I set $\gamma = 6$. I present details on the solution method in Appendix B. The main difference is that, due to the short selling costs, the price is not linear in Y_t and it is not possible to obtain an analytical expression. In the same way, V_t must be obtained numerically.

Figure 1.7 depicts the deviations of price from equilibrium, $P_t - V_t$, in the presence of short selling costs. The left panel presents the results for a simulation time of $T = 100$ periods, at whose end the fluctuations are still material. I thus simulate the system for $T = 200$ periods, showing the results in the right panel of the figure. The longer simulation shows that the price deviations are progressively dampened. Comparing these results with the price dynamics of Figure 1.6, short selling costs make deviations from equilibrium more persistent. Fluctuations take more time to dampen. This fact is in line with the results of Anufriev and Tuinstra (2013) and with the framework of Miller (1977).

Finally, I compare the effect of an over-pricing and an under-pricing shock. The goal is to to see whether this framework accounts for the different impact on the ECT, in line with my empirical analysis. I present my results in Figure 1.8. In both cases I impose a supply shock such that the fundamental value changes by 10%. The blue line corresponds to an underpricing shock, due to a sudden positive increase in supply, as in Figure 1.7. The orange line corresponds instead to a decrease in supply, that delivers an overpricing shock, as the price jumps above the fundamental value. We can think about this setting as if we had the two legs of an anomaly portfolios, which receive an opposite shock. Given an equivalent change in fundamental, the short leg (overpricing shock) has a larger ECT on impact, which requires more time to dampen, thus being more persistent. Therefore,



(a) $T = 100$ **(b)** $T = 200$

Figure 1.7: The figure depicts the price deviations from the fundamental value after a positive supply shock at time $t = 1$. The model includes short selling logistic costs. Panel (a) considers a simulation duration of $T = 100$ periods, while Panel (b) has $T = 200$ periods.

accounting for short-selling costs brings the model in line with the empirical framework, accounting for the differential persistence that characterizes the long and short legs of anomaly portfolios.

1.7 Conclusion

Large cross-sections of long-short anomaly portfolios predict the market excess return (Dong et al., 2022). This chapter investigates why. I decompose the *prices* of the long and short legs of the anomalies into permanent and transitory components. By using dimension reduction techniques, I show that the latter predict the market excess *return*. This result holds at multiple horizons and is mostly driven by the long components of the anomaly portfolios. I interpret these findings through an asymmetric limits of arbitrage model.

These results highlight the importance of modeling asset prices together with returns. They also contribute to a new literature that links the cross-sectional and time-series predictability in the stock market. An extension of this work could be to develop theoretical

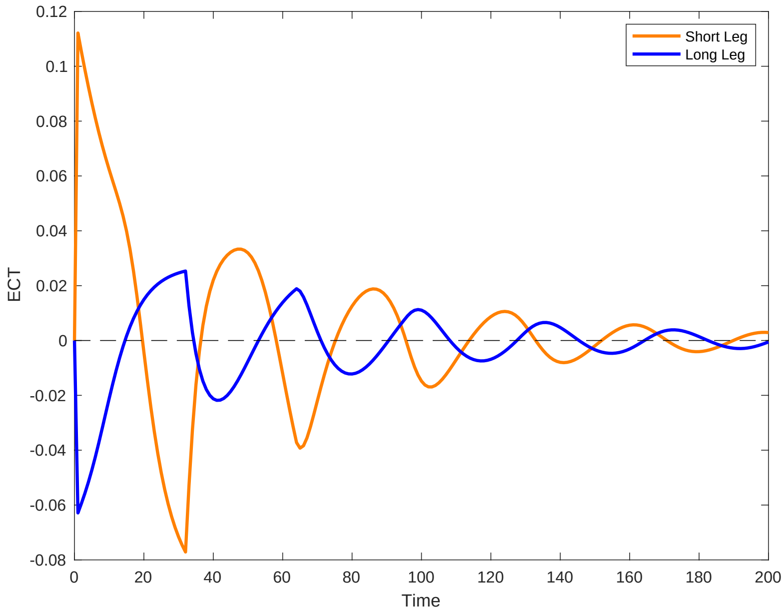


Figure 1.8: The figure depicts the price deviations from the fundamental value after a positive (blue line) and negative (orange line) supply shock at time $t = 1$. The shocks are chosen such that the fundamental value changes by 10%.

models to study the price dynamics in closed form.

A The Econometric Framework

Proposition 1 (Long vs Short Legs). *Under Assumption 1, the ratio between the Long and Short portfolios standardized regression coefficients is*

$$\left|\frac{\tilde{\beta}_L}{\tilde{\beta}_S}\right| = \sqrt{\frac{1-\rho_L}{1-\rho_S}}\sqrt{\frac{1+\rho_S}{1+\rho_L}} > 1. \tag{1.11}$$

Moreover, the ratio between the mean squared forecast errors is

$$\frac{\mathbb{E}\left[\epsilon^2_{L,t+1}\right]}{\mathbb{E}\left[\epsilon^2_{S,t+1}\right]} < 1. \tag{1.12}$$

Proof. The standardized regression coefficient is

$$\tilde{\beta}_l = \frac{Cov\left(r_{M,t,t+1}, u_{l,t}\right)}{\sqrt{V\left(u_{l,t}\right)}}.$$

Calculations show

$$\frac{Cov\left(r_{M,t,t+1}, u_{l,t}\right)}{\sqrt{V\left(u_{l,t}\right)}} = \frac{Cov\left(f_{t,t+1} + \frac{1}{2}\left(\rho_L - 1\right)u_{L,t} + \frac{1}{2}\left(\rho_S - 1\right)u_{S,t} + \eta_{t+1}, u_{l,t}\right)}{\sqrt{V\left(u_{l,t}\right)}} =$$
$$= \frac{1}{2}\left(\rho_l - 1\right)\sqrt{V\left(u_{l,t}\right)}.$$

Since we have an AR(1) process, $V\left(u_l\right) = V\left(\eta_l\right)/\left(1-\rho_l^2\right)$. I assume $V\left(\eta_L\right) = V\left(\eta_S\right)$. Because of asymmetric limits of arbitrage, $\rho_S > \rho_L$, and

$$\frac{\tilde{\beta}_L}{\tilde{\beta}_S} = \frac{\left(\rho_L - 1\right)\sqrt{V\left(u_L\right)}}{\left(\rho_S - 1\right)\sqrt{V\left(u_S\right)}} = \frac{1-\rho_L}{1-\rho_S}\sqrt{\frac{1-\rho_S^2}{1-\rho_L^2}} = \sqrt{\frac{1-\rho_L}{1-\rho_S}}\sqrt{\frac{1+\rho_S}{1+\rho_L}}$$
$$> \sqrt{\frac{1-\rho_S}{1-\rho_S}}\sqrt{\frac{1+\rho_L}{1+\rho_L}} = 1.$$

Now, concerning the MSFE,

$$\begin{aligned}\epsilon_{l,t+1} &= r_{M,t,t+1} - \alpha_l - \beta_l u_{l,t} \\ &= f_{t,t+1} + \frac{1}{2}\left(\rho_L - 1\right) u_{L,t} + \frac{1}{2}\left(\rho_S - 1\right) u_{S,t} + \eta_{t+1} - \alpha_l - \frac{1}{2}\left(\rho_l - 1\right) u_{l,t} \\ &= f_{t,t+1} + \frac{1}{2}\left(\rho_{-l} - 1\right) u_{-l,t} + \eta_{t+1} - \alpha_l,\end{aligned}$$

and since $\left(1-\rho_L\right)^2 > \left(1-\rho_S\right)^2$,

$$\mathbb{E}\left[\epsilon^2_{S,t+1}\right] > \mathbb{E}\left[\epsilon^2_{L,t+1}\right].$$

□

Proposition 2 (ECT vs Returns)**.** *Under $\rho_l < 1$, the ratio between the standardized regression coefficients is*

$$\left|\frac{\tilde{\beta}_{r,l}}{\tilde{\beta}_l}\right| < 1. \tag{1.14}$$

Moreover,

$$\frac{\mathbb{E}\left[\epsilon^2_{r,l,t+1}\right]}{\mathbb{E}\left[\epsilon^2_{l,t+1}\right]} > 1. \tag{1.15}$$

Proof. Consider the regressions

$$\begin{aligned}r_{M,t,t+1} &= \alpha_{r,l} + \beta_{r,l} r_{l,t-1,t} + \epsilon_{r,l,t+1}, \\ r_{M,t,t+1} &= \alpha_l + \beta_l u_{l,t} + \epsilon_{l,t+1},\end{aligned}$$

and let us compute the beta coefficients:

$$\begin{aligned}\beta_{r,l} &= \frac{Cov\left(r_{M,t,t+1}, r_{l,t-1,t}\right)}{\sqrt{V\left(r_l\right)}} = \frac{\rho_l\left(1-\rho_l\right)^2 V\left(u_l\right)}{2\sqrt{V\left(r_l\right)}} > 0, \\ \beta_{u,l} &= \frac{Cov\left(r_{M,t,t+1}, u_{l,t}\right)}{\sqrt{V\left(u_l\right)}} = \frac{\left(\rho_l - 1\right) V\left(u_l\right)}{2\sqrt{V\left(u_l\right)}} < 0.\end{aligned}$$

The correlation between the trends is zero, as they are not at the same time. Let us look at the ratio between the coefficients:

$$\begin{aligned}\frac{\beta_{r,l}}{\beta_{u,l}} &= \rho_l\left(\rho_l-1\right)\sqrt{\frac{V\left(u_l\right)}{V\left(r_l\right)}} = \rho_l\left(\rho_l-1\right)\sqrt{\frac{V\left(u_l\right)}{V\left(f_l\right)+\left(\rho_l-1\right)^2V\left(u_l\right)+V\left(\eta_l\right)}} \\ &= \frac{\rho_l\left(\rho_l-1\right)}{\left(1-\rho_l\right)}\sqrt{\frac{\left(\rho_l-1\right)^2V\left(u_l\right)}{V\left(f_l\right)+\left(\rho_l-1\right)^2V\left(u_l\right)+V\left(\eta_l\right)}} = -\rho_l\sqrt{\frac{\left(\rho_l-1\right)^2V\left(u_l\right)}{V\left(f_l\right)+\left(\rho_l-1\right)^2V\left(u_l\right)+V\left(\eta_l\right)}}.\end{aligned}$$

Since $\rho_l < 1$ and the argument of the square root is smaller or equal to unity,

$$\left|\frac{\beta_{r,l}}{\beta_{u,l}}\right| < 1, \tag{38}$$

i.e. the ECT predicts better. Notice there are also some predictions on the sign of the coefficients. Now, concerning the MSFE,

$$\begin{aligned}\epsilon_{r,l,t+1} &= r_{M,t,t+1}-\alpha_{r,l}-\beta_{r,l}r_{l,t-1,t} \\ &= f_{t,t+1}+\frac{1}{2}\left(\rho_L-1\right)u_{L,t}+\frac{1}{2}\left(\rho_S-1\right)u_{S,t}+\eta_{t+1}+ \\ &-\alpha_{r,l}-\frac{\rho_l\left(1-\rho_l\right)^2V\left(u_l\right)}{2V\left(r_l\right)}\left(f_{l,t-1,t}+\frac{1}{2}\left(\rho_l-1\right)u_{l,t-1}-\eta_t\right).\end{aligned}$$

□

Proposition 3 (Multi-horizon Regressions)**.** *Under cointegration ($\rho_l < 1$), the ratio between the standardized regression coefficients is*

$$\left|\frac{\tilde{\beta}_l^{k+n}}{\tilde{\beta}_l^k}\right| > 1, \quad n \geq 1. \tag{1.17}$$

Moreover,

$$\frac{\mathbb{E}\left[\epsilon^2_{l,t+k+n}\right]}{\mathbb{E}\left[\epsilon^2_{l,t+k}\right]} < 1. \tag{1.18}$$

Proof. Consider the regression

$$r_{M,t,t+k} = \alpha_l^k + \beta_l^k u_{l,t} + \epsilon_{l,t+k}.$$

The standardized regression coefficient is

$$\tilde{\beta}_l^k = \frac{Cov\left(r_{M,t,t+k}, u_{l,t}\right)}{\sqrt{V\left(u_{l,t}\right)}}.$$

Calculations show

$$\frac{Cov\left(r_{M,t,t+k}, u_{l,t}\right)}{\sqrt{V\left(u_{l,t}\right)}} = \frac{Cov\left(f_{t,t+k} + \frac{1}{2}\left(\rho_L^k - 1\right)u_{L,t} + \frac{1}{2}\left(\rho_S^k - 1\right)u_{S,t} + \eta_{t+k}, u_{l,t}\right)}{\sqrt{V\left(u_{l,t}\right)}} =$$
$$= \frac{1}{2}\left(\rho_l^k - 1\right)\sqrt{V\left(u_{l,t}\right)}.$$

We can show that $\left|\frac{\tilde{\beta}_l^{k+n}}{\tilde{\beta}_l^k}\right| > 1$. Indeed

$$\left|\frac{\tilde{\beta}_l^{k+n}}{\tilde{\beta}_l^k}\right| = \frac{(1-\rho_l^{k+n})}{(1-\rho_l^k)} > 1,$$

as $\rho_l < 1$ and $\rho_l^{k+n} < \rho_l^k$. The second part of the Proposition is a consequence of writing

$$\epsilon_{l,t+k} = r_{M,t,t+k} - \alpha_l^k - \beta_l^k u_{l,t}$$

and substituting the expression for the coefficients, as in the previous Propositions. □

B The Model

This section shows the details of the proofs of the model.

B.1 Attentive Investors

Let K_t^i be the demand for the risky asset by individual investor i, with risk-aversion coefficient ϕ_i. Each investor has constant absolute risk aversion (CARA) utility, so that the maximization problem to be solved is

$$\max_{C_t^i, K_t^i} \mathbb{E}_t \left(W_{t+1}^i\right) - \frac{\phi_i}{2} \mathbb{V}_t \left(W_{t+1}^i\right), \tag{39}$$

subject to the budget constraint

$$W_{t+1}^i = \left(W_t^i - C_t^i\right) R + K_t^i \left(P_{t+1} - RP_t + X_{t+1}\right). \tag{40}$$

Expectations are homogeneous. By considering the first order condition with respect to K_t^i, the optimal demand is

$$K_t^i = \frac{\mathbb{E}_t \left[P_{t+1} + X_{t+1}\right] - RP_t}{\phi_i \mathbb{V}_t \left[P_{t+1} + X_{t+1}\right]}. \tag{41}$$

The attentive investors' aggregate demand is obtained by integrating,

$$K_t = \int di K_t^i = (1-q) \frac{\mathbb{E}_t \left[P_{t+1} + X_{t+1}\right] - RP_t}{\phi \mathbb{V}_t \left[P_{t+1} + X_{t+1}\right]}, \tag{42}$$

in which ϕ is the harmonic mean of the risk-aversion coefficients[9].

[9]The harmonic mean is defined as $\phi = \left(\int di \left(\phi_i\right)^{-1}\right)^{-1}$.

B.2 Inattentive Investors

Let D_t^i be the demand for the risky asset of the i-th inattentive investor. The risk-aversion coefficient is θ_i, and each investor agent has CARA utility. The maximization problem is

$$\max_{C_t^i, D_t^i} \mathbb{E}_t \left(W_{t+k}^i\right) - \frac{\theta_i}{2} \mathbb{V}_t \left(W_{t+k}^i\right), \tag{43}$$

subject to the budget constraint

$$W_{t+k}^i = \left(W_t^i - C_t^i\right) R^k + D_t^i \left(P_{t+k} + \sum_{n=1}^{k} R^{k-n} X_{t+n} - R^k P_t\right). \tag{44}$$

Solving for D_t^i and integrating over the inattentive investors yield the aggregate demand

$$D_t = \frac{q}{k} \frac{\mathbb{E}_t \left[P_{t+k} + \sum_{n=1}^{k} R^{k-n} X_{t+n}\right] - R^k P_t}{\theta \mathbb{V}_t \left[P_{t+k} + \sum_{n=1}^{k} R^{k-n} X_{t+n}\right]}, \tag{45}$$

with θ being the harmonic mean of the risk-aversion coefficients.

B.3 The solution

At time t, the inattentive investors who were active in the previous $k-1$ periods are locked in their position. The market clearing condition reads

$$K_t + D_t = Z_t - D_{t-1} - ... - D_{t-k+1}. \tag{46}$$

The relevant state vector is $Y_t = (Z_t, X_t, H_t)'$, with $H_t = (D_{t-1}, ..., D_{t-k+1})'$. It can be proved that Y_t follows the process

$$Y_t = AY_{t-1} + B\epsilon_t. \tag{47}$$

In this setting, the demand and price are linear in Y_t: $D_t = aY_t$, $K_t = bY_t$ and $P_t = cY_t$. See the appendix in Duffie (2010) for the expressions. I find c with MATLAB. I specialize my results to the case

$$\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \Sigma^{1/2} = \begin{pmatrix} 0 & 0 \\ 0 & \sqrt{0.1} \end{pmatrix}. \tag{48}$$

Differently from Duffie (2010), I express my results in terms of deviations from the equilibrium price V_t. I find V_t following the intuition in Anufriev and Tuinstra (2013), that is as the steady state. This approach is appropriate given the special form of the Λ and $\Sigma^{1/2}$ matrices. In particular, I impose

$$K + kD = Z, \tag{49}$$

with

$$K = \frac{(1-q)(1-R)}{\phi\sigma_1^2} P, \tag{50}$$

$$kD = \frac{q\left(1-R^k\right)}{\theta\sigma_k^2} q. \tag{51}$$

The variances are

$$\sigma_1^2 = \mathbb{V}(X) = \sigma_X^2, \tag{52}$$

$$\sigma_k^2 = \mathbb{V}\left(\sum_{i=1}^{k} R^{k-i} X\right) = \sum_{i=1}^{k} \left(R^2\right)^{k-i} \sigma_X^2 = \frac{1-R^{2k}}{1-R^2}\sigma_X^2. \tag{53}$$

Using the market clearing condition, we can express the fundamental value as

$$V = \frac{\phi Z}{(1-q)(1-R)/\sigma_1^2 + q\left(1-R^k\right)/\sigma_k^2}. \tag{54}$$

B.4 Short Selling Costs

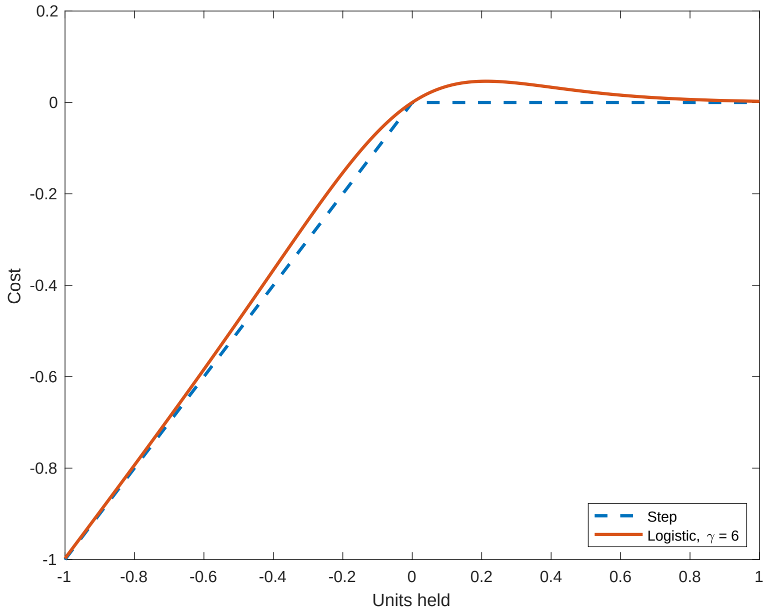


Figure 9: This figure depicts the short selling costs under a step function (dashed line) and a logistic function (solid line).

C Additional Results

Table V
R^2_{OS} Statistics - Economic Restrictions

The table reports Campbell and Thompson (2008) out-of-sample R^2 (R^2_{OS}) in percent for market excess return forecasts at horizons of one month, one quarter and one year. The model forecasts are forced to be greater than zero, to have a non-negative equity premium. The out-of-sample period is 2000:01 to 2017:12. U-PCA forecasts are based on the first principal component of the cointegrated error correction terms of the portfolios. U-ENet forecasts are built from elastic net predictive regressions that include all the cointegrated error correction terms (returns) of the portfolios. L&S, L, S refer to the cross-section of long and short, long, short portfolios. *, ** and *** indicate significance at the 10%, 5% and 1% level respectively for the positive R^2_{OS}, based on the Clark and West (2007) test.

	L		S		L&S	
	U-PCA	U-ENet	U-PCA	U-ENet	U-PCA	U-ENet
$R_{M,t,t+1}$	0.76**	0.42	-0.63	-1.65	0.60**	-2.00
	(1.70)	(1.05)	(-1.24)	(-0.65)	(1.90)	(-0.9)
$R_{M,t,t+3}$	1.37**	−1.05	-2.38	-5.94	1.21***	-7.34
	(1.88)	(−0.02)	(-1.60)	(-1.19)	(2.33)	(-1.69)
$R_{M,t,t+12}$	2.74***	2.09*	-5.04	-8.70	1.36**	−1.49
	(2.26)	(1.61)	(-3.40)	(-1.42)	(1.67)	(1.18)

Table VI
R^2_{OS} Statistics - Alternative Dataset

The table reports Campbell and Thompson (2008) out-of-sample R^2 (R^2_{OS}) in percent for market excess return forecasts at horizons of one month, one quarter and one year. The out-of-sample period is 2000:01 to 2017:12. U-PCA forecasts are based on the first principal component of the cointegrated error correction terms of the portfolios. U-ENet forecasts are built from elastic net predictive regressions that include all the cointegrated error correction terms (returns) of the portfolios. The anomaly portfolios are the 154 anomalies from Jensen et al. (2023). *, ** and *** indicate significance at the 10%, 5% and 1% level respectively for the positive R^2_{OS}, based on the Clark and West (2007) test.

	U-PCA	U-ENet
$R_{M,t,t+1}$	1.74**	−2.64
	(2.00)	(−1.46)
$R_{M,t,t+3}$	6.32***	−0.08
	(4.02)	(−0.93)
$R_{M,t,t+12}$	9.89***	5.86***
	(5.35)	(3.97)

References

Anufriev, M. and Tuinstra, J. (2013). The impact of short-selling constraints on financial market stability in a heterogeneous agents model. *Journal of Economic Dynamics and Control*, 37(8):1523–1543.

Baba Yara, F., Boons, M., and Tamoni, A. (2020). New and old sorts: Implications for asset pricing. *Martijn and Tamoni, Andrea, New and Old Sorts: Implications for Asset Pricing (January 31, 2020).*

Bai, J. (2004). Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics*, 122(1):137–183.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150.

Banerjee, A. and Marcellino, M. (2009). Factor-augmented error correction models. *The methodology and practice of econometrics?a festschrift for David Hendry*, pages 227–254.

Banerjee, A., Marcellino, M., and Masten, I. (2014). Forecasting with factor-augmented error correction models. *International Journal of Forecasting*, 30(3):589–612.

Bansal, R., Dittmar, R., and Kiku, D. (2009). Cointegration and consumption risks in asset returns. *The Review of Financial Studies*, 22(3):1343–1375.

Bansal, R. and Kiku, D. (2011). Cointegration and long-run asset allocation. *Journal of Business & Economic Statistics*, 29(1):161–173.

Bianchi, D., Rubesam, A., and Tamoni, A. (2023). It takes two to tango: Economic theory and model uncertainty for equity premium prediction. *Available at SSRN 4513241*.

Cakici, N., Fieberg, C., Metko, D., and Zaremba, A. (2024). Do anomalies really predict market returns? new data and new evidence. *Review of Finance*, 28(1):1–44.

Campbell, J. Y. and Shiller, R. J. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *The review of financial studies*, 1(3):195–228.

Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531.

Cho, T. and Polk, C. (2020). Putting the price in asset pricing. *Available at SSRN 3499681*.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics*, 138(1):291–311.

Cohen, R. B., Polk, C., and Vuolteenaho, T. (2009). The price is (almost) right. *The Journal of Finance*, 64(6):2739–2782.

Dong, X., Li, Y., Rapach, D. E., and Zhou, G. (2022). Anomalies and the expected market return. *The Journal of Finance*, 77(1):639–681.

Duffie, D. (2010). Presidential address: Asset price dynamics with slow-moving capital. *The Journal of finance*, 65(4):1237–1267.

Fama, E. F. and French, K. R. (1988). Permanent and temporary components of stock prices. *Journal of political Economy*, 96(2):246–273.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Favero, C. A., Melone, A., and Tamoni, A. (2021). Factor models with drifting prices. *Available at SSRN 3418352*.

Flynn, C. J., Hurvich, C. M., and Simonoff, J. S. (2013). Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. *Journal of the American Statistical Association*, 108(503):1031–1043.

Goyal, A., Welch, I., and Zafirov, A. (2023). A comprehensive look at the empirical performance of equity premium prediction ii. *Swiss Finance Institute Research Paper*, (21-85).

Gromb, D. and Vayanos, D. (2010). Limits of arbitrage. *Annu. Rev. Financ. Econ.*, 2(1):251–275.

Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518.

Lettau, M. and Ludvigson, S. (2001). Consumption, aggregate wealth, and expected stock returns. *the Journal of Finance*, 56(3):815–849.

Lettau, M. and Ludvigson, S. C. (2005). Expected returns and expected dividend growth. *Journal of Financial Economics*, 76(3):583–626.

Martin, I. W. and Nagel, S. (2021). Market efficiency in the age of big data. *Journal of Financial Economics*.

McCracken, M. W. and Valente, G. (2018). Asymptotic inference for performance fees and the predictability of asset returns. *Journal of Business & Economic Statistics*, 36(3):426–437.

Miller, E. M. (1977). Risk, uncertainty, and divergence of opinion. *The Journal of finance*, 32(4):1151–1168.

Newey, W. K. and West, K. D. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix.

Shleifer, A. and Vishny, R. W. (1997). The limits of arbitrage. *The Journal of finance*, 52(1):35–55.

van Binsbergen, J. H., Boons, M., Opp, C. C., and Tamoni, A. (2021). Dynamic asset (mis) pricing: Build-up vs. resolution anomalies. *Resolution Anomalies (February 12, 2021)*.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

Chapter 2

News Headlines

This chapter is based on the paper "News Headlines", which is joint work with Massimiliano Marcellino.

2.1 Introduction

Text analysis has become pervasive. It unleashes a wealth of new information, which can be used by researchers and policy makers to shed more light on the workings of the economy. The type of text analyzed in the literature ranges from institutional documents (Hansen et al., 2018) to earnings conference calls (Hassan et al., 2019), from social media posts (Bianchi et al., 2021) to business news (Bybee et al., 2021). Text is used to construct uncertainty indicators (Baker et al., 2016; Manela and Moreira, 2017), measure the state of the economy (Bybee et al., 2021) and forecast economic variables (Kelly et al., 2021). However, text data comes at a cost: it is ultra-high dimensional. Developing methods to handle such complexity, balancing the information content of text and its interpretation, is key. In this chapter, we focus on a specific type of text: the headlines of news articles. Our main contribution lies in exploring the relevance of headlines for economic applications.

Headlines are a summary of the whole text, and they may be interpreted as a signal extracted from it. Empirical evidence shows that people focus on headlines while consuming and sharing information (Gabielkov et al., 2016), which makes them relevant to model beliefs. On the other hand, the full text contains additional content by definition, which may be crucial to measure the state of the economy and predict its future trajectory. Since headlines are designed to catch the readers' attention, they may also provide a biased and distorted version of the true article's content. Whether relying solely on headlines in place of the complete text might hinder economic analysis, such as in the context of forecasting, is an empirical question.

We provide some graphical intuition in Figure 2.1. The figure shows two standardized measures of sentiment we extracted respectively from the headlines (orange line) and the full text (blue line) of more than $400,000$ *Wall Street Journal* articles covering the period from $2008:01$ to $2022:04$. The two series closely track each other, with the headlines one assuming slightly larger absolute values during peak and troughs.

We compare the two sentiment series in an out-of-sample analysis. We use them separately as predictors for 12 macroeconomic indicators, besides the top 5 principal components from a large macro-financial panel (Stock and Watson, 2012). We consider forecasting horizons from 1 to 12 months. Interestingly, the relative mean square forecast error (RMSE) ratios for headlines and full texts are mostly close to 1, and none are significantly different from 1, indicating that the forecasts from the two sources are equally accurate.

We repeat our analysis using a Hurdle Distributed Multinomial Regression (HDMR) model (Kelly et al., 2021), a state-of-the-art machine learning method for text-based forecasting. We compare the RMSE obtained using only the headlines or the full text in the out-of-sample exercise. Once again, the RMSE ratios are close to 1, and only a handful of ratios are significantly different from 1.

As a robustness check, we restrict our sample to the front page articles. Our results are even stronger. The only statistically significant RMSE ratios, e.g. for the S&P500 index,

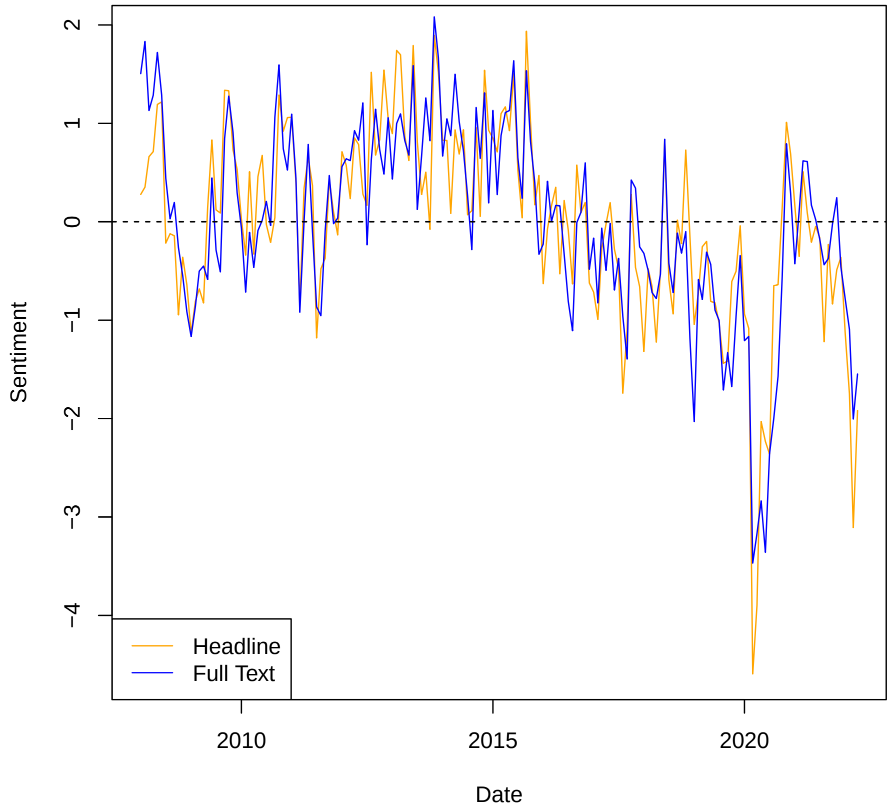


Figure 2.1: The picture depicts sentiment measures extracted from the headlines (orange line) and the full text (blue line) of the *Wall Street Journal*, covering the period from 2008 : 01 to 2022 : 04.

are lower than 1, indicating that headline-based forecasts are more accurate. Moreover, the front page forecasts are comparable to the ones built from the full set of articles. We also split our out-of-sample evaluation window between pre and post Covid, and our results are unaffected. Overall, our findings suggest that the additional content in the full text does not add value to forecasting economic variables with respect to using headlines only, despite the full text containing potentially much more information.

We interpret our findings through an accuracy-clarity trade-off, in the spirit of Myatt and Wallace (2012). Our framework combines elements from the rational inattention (Sims, 2003) and beauty contest (Keynes, 1936) literature. We build a model where the agents need to forecast an economic indicator using the information sent by a news producer. These readers pay an attention cost for processing the information. Moreover, the agents are incentivized to focus on what others are reading, as deviating from the general consensus involves a penalty (Morris and Shin, 2002).

In our model, the news producer sends two signals for the underlying economic variable: the headline and the full text. The headline is less accurate, i.e. it has more noise than the full text. The readers receive a noisy version of the two signals. The headline signal is clearer, i.e. it displays less noise on the receiver side. This modeling assumption mirrors the fact that while headlines contain less information than the full text, they are easier to interpret. Under simplified assumption, we show that if enough agents have a high attention cost, there is an equilibrium where every agent only uses the signal from the headlines. Despite having a more accurate signal, the full text is less relevant for forecasting.

The remainder of the chapter is organized as follows. Section 2.2 describes our data. Section 2.3 presents the methodology used in our analysis. Section 2.4 presents the empirical results and discusses their implications. Section 2.5 describes a theoretical model to frame our results. Section 2.6 concludes the chapter.

2.2 Data

This section describes the data we use.

Wall Street Journal We collected all the *Wall Street Journal* (WSJ) articles from Factiva for the $2008:01$ to $2022:04$ period, at the daily frequency. We exclude articles unrelated to economics, such as sports articles. Our final dataset includes more than $400,000$ distinct articles. In a robustness exercise, we will only consider the front page articles, with an average of 166 items per month.

Table I
Summary Statistics

This table reports the summary statistics for 12 macro-financial indicators we use as targets. The sample goes from 2008 : 01 to 2021 : 12, at the monthly frequency. We transform the series as in Stock and Watson (2012): we take first differences of logarithms (growth rates) for real quantity variables, first differences for nominal interest rates, and second differences of logarithms (changes in rates of inflation) for price series.

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max	N
IP: total	0.616	2.623	-16.153	-0.702	0.845	2.191	8.314	230
Emp: total	0.353	0.755	-2.671	-0.064	0.474	0.899	1.801	230
U: all	0.012	0.150	-0.400	-0.100	0.000	0.100	0.500	230
HStarts: total	7.281	0.247	6.190	7.149	7.320	7.420	7.729	230
PMI	51.274	5.102	32.900	49.000	51.650	54.800	61.400	230
CPI-ALL	0.001	1.165	-4.467	-0.550	-0.002	0.647	4.302	230
Real AHE: goods	0.222	1.131	-4.211	-0.357	0.177	0.757	5.782	230
FedFunds	-0.036	0.206	-0.960	-0.100	0.000	0.070	0.530	230
M1	-0.008	2.505	-10.512	-1.351	-0.044	1.394	7.485	230
Ex rate: avg	-0.164	6.863	-19.402	-4.380	0.515	4.545	21.450	230
SP 500	1.456	14.873	-91.237	-4.988	2.617	10.643	42.821	230
Consumer expect	-0.152	5.003	-14.400	-3.075	-0.400	2.575	22.500	230

Macro-financial Variables We have a panel of 92 monthly indicators from Stock and Watson (2012), which we use to compute principal components, together with 12 headline indicators listed in Table I, which will be our targets. The sample is monthly, from 2008 : 01 to 2021 : 12. We apply standard transformations, and treat outliers as in Stock and Watson (2012). In particular, we identify outliers as observations with absolute median deviation larger than 6 times the interquartile range, and we replace them with the median of the previous 5 observations. We report some summary statistics in Table I.

2.3 Empirical Framework

In this section we describe our empirical approach. We will conduct our analysis both in-sample and out-of-sample. We first compare sentiment measures from the headline versus the full body of news articles, which provides the most intuitive comparison. We later use a state-of-the-art machine learning approach, to see whether our findings carry out to this case.

2.3.1 Measuring Sentiment

As we already mentioned, the goal of our analysis is to extract information from two different types of text, the headline versus the full-body. The simplest approach is to start from a simple sentiment measure, which is simple to compute and analyze. In this section, we describe how we compute our sentiment measure. We implement the method described below by using the R package `sentimentr`.

We compute sentiment following the lexicon approach, a commonly used method in sentiment analysis. It involves using a pre-defined set of words with assigned positive, negative, or neutral sentiment scores to compute a sentiment measure for a given text. We first select an appropriate lexicon that matches the context of the text being analyzed, i.e. a dictionary. Since we are dealing with economic and business news, we cannot use a general purpose dictionary, such as the Harvard General Inquirer. The reason is that some technical economic words have a different meaning in non-financial language[1]. Following Shapiro et al. (2020), we combine a general purpose dictionary with the Loughran and McDonald (2011) updated lexicon, which is specific to the domain of economics and finance.

After choosing a dictionary, the next step is to preprocess the text being analyzed by removing stopwords, punctuations, and other irrelevant information. After preprocessing, the corresponding sentiment score for each word is retrieved. The sentiment scores of all words in the text are then aggregated to obtain a sentiment measure for the entire text. We later average the sentiment over all the text produced in a given period (such as a month), and we down-weight the zeros. In this way, we make sure neutral sentences do not have too much of a strong influence on our score.

[1]For example, the word “liability” would be misclassified as negative with a general purpose dictionary, while it does not have a negative implication in finance.

Moreover, we take into account the effect of valence shifters on the sentiment of a text. Valence shifters are words that can change the polarity of the sentiment of a sentence or phrase. For example, the word "not" is a valence shifter that can change a positive sentiment to a negative sentiment. Other valence shifters include intensifiers such as "very" and "extremely" that can increase the intensity of a sentiment, as well as negations such as "never" and "no one" that can negate the sentiment of a sentence. We explicitly account for such shifters in our sentiment algorithm.

The procedure described above extracts information from text, via a simple natural language processing (NLP) algorithm. Shapiro et al. (2020) show that such a measure is highly correlated with the consumer sentiment series coming from the University of Michigan survey. Also our measure is highly correlated to consumer sentiment, with a correlation coefficient close to 70% over the whole sample. This fact leads us to call this series "sentiment", giving to it a behavioral connotation.

We follow the steps above to compute a measure of sentiment first from the headlines alone, and then from the full body of the articles (excluding the headlines). We aggregate the sentiment to obtain a monthly series. Notice that we obtain a "level" indicator, that looks at the overall difference between positive and negative counts. One might also try to compute a "dispersion" measure, so that months with highly varying (in tone) news that average to zero (e.g. several extremely positive and negative news that cancel out) are treated differently from months with only neutral news. Adding a measure of news dispersion to our setting does not affect our results and conclusions. However, it may be a promising avenue to include such a dimension to text-based forecasting models.

2.3.2 Out-of-Sample Analysis

We run an out-of-sample analysis to compare the information contained in the headlines and full text. We test whether including the additional information in the body leads to more accurate forecasts. We run regressions in the form

$$y_{t+\tau} = \boldsymbol{\beta}'\boldsymbol{x}_t + \boldsymbol{\gamma}\boldsymbol{z}_t^s + \epsilon_t. \tag{2.1}$$

$y_{t+\tau}$ is a headline macro-financial indicator at time $t+\tau$. The vector $\boldsymbol{x}_t = (1, PC_1, ..., PC_5)'$ contains a constant and the first 5 principal components, extracted from a large panel of macro-financial variables as in Stock and Watson (2012). $\boldsymbol{z}_t^s$ collects the variables which contain the information extracted from text of type s, with $s = h$ if only the headlines are used to construct such measure, and $s = f$ if the full body is used.

In the simplest case, $\boldsymbol{z}_t^s$ only contains the sentiment extracted from s as explained in Section 2.3.1. We also compute $\boldsymbol{z}_t^s$ via a state-of-the-art machine learning model, the Hurdle Distributed Multinomial Regression (HDMR) developed by Kelly et al. (2021). We describe this method shortly in Section 2.3.3.

We obtain forecasts of $y_{t+\tau}$ at different horizons τ, ranging from 1 to 12 months ahead. We compute the Root Mean Squared Errors (RMSE) for each horizon τ and type of text s, to later compute their ratio $RMSE_\tau^h/RMSE_\tau^b$. A ratio smaller than 1 means that the signal extracted from the headlines forecasts better than the full body. We compare the OOS performance using the Diebold and Mariano (2002) approach, with the Harvey et al. (1997) correction.

2.3.3 Hurdle Distributed Multinomial Regression

We briefly describe the machine learning model developed by Kelly et al. (2021), the Hurdle Distributed Multinomial Regression (HDMR). It allows to forecast economic indicators starting from text data, using a two-part model which adapts Heckman (1979) to a high-dimensional setting. HDMR combines a selection equation, which models the text producer's choice of whether or not to include a particular phrase, with a positive counts model, which models the number of times that phrase is repeated. The advantage of such model is that it provides an accurate description of word counts in text data and can be distributed across parallel computing units, which makes its estimation feasible. Moreover, rather than extracting an independent sentiment indicator that is later used to forecast a variable, this approach directly models the text information that is useful to predict a given indicator. In this sense, it could be seen as a similar approach (in spirit) to Huang et al. (2015).

The input of the model is a count matrix, whose entries c_{tj} contain the number of times phrase j is included in document t. t can be either a single document or, as in our case, all the articles produced in a time period (e.g. all the WSJ articles in a month). For our analysis, phrase j will be either a monogram (one word) or a bigram (two contiguous words). Suppose we are interested in predicting a variable y using our text. While matrix $\{c_{tj}\}$ is ultra high-dimensional, Kelly et al. (2021) prove that the text content that is useful for predicting y is summarized by two low dimension sufficient statistics, z_t^0 and z_t^+. z_t^0 contains all the information which is useful to predict y from the selection of phrases in the text. z_t^+ summarize all the information from repeating words in text t. These two variables are collected in the vector $\boldsymbol{z}_t$ and used as predictors in an out-of-sample forecasting exercise of y.

Our empirical strategy consists in estimating $\boldsymbol{z}_t^s$ using only phrase counts from the headlines ($s = h$) or the full text ($s = f$) of the articles. In order to construct the count matrix used as an input for HDMR, we apply standard text transformations, such as removing case, common stopwords (e.g. "the") and Porter Stemming to our sample. We count the number of monograms and bigrams, going from the raw text to a matrix of counts. We later estimate Equation (2.1) using one $\boldsymbol{z}_t^s$ vector at a time, and we compare the RMSE we obtain in each case as described in Section 2.3.2.

Finally, notice that we count the words (or n-grams) contained in the full text or the headlines, but later aggregate our counts at the monthly level. We do so in line with the literature. We do not specialize the analysis at the article level, as we would have too much noise and not enough repetitions.

2.4 Results

This section presents our main results. We start by comparing the sentiment measures in-sample, which provides a simple graphical intuition behind our findings. We later conduct an out-of-sample analysis, using both the sentiment measures and the more complicated HDMR model.

2.4.1 Sentiment

We compute two different sentiment measures, starting from the set of all the headlines and all the full body of the news articles in the *Wall Street Journal*. We plot them in Figure 2.1. The figure depicts the time series of the headline (orange line) and full body (blue line) standardized sentiment. The two series closely track each other, with the headlines one assuming slightly larger absolute values during peak and troughs. The correlation coefficient is 0.87. The main difference is that the headline series is more volatile.

The previous analysis shows that the two series are highly correlated. Hence, we would expect them to have a similar accuracy, if they were to be used as predictors. We conduct an out-of-sample forecasting exercise as in Section 2.3.2, using one sentiment measure at a time (together with the 5 principal components from Stock and Watson (2012)) to predict 12 different macro-financial variables. We use forecasting horizons of 1, 3 and 12 months. The out-of-sample window is 2015 to 2021 for each horizon, and we use rolling windows with monthly observations, as in Kelly et al. (2021).[2] Table II reports our results on the right panel.

The right panel of Table II confirms our intuition from Figure 2.1. Most of the RMSE

[2]Even though we can aggregate news data at a higher frequency, e.g. daily or weekly, we use a monthly frequency, in line with Kelly et al. (2021). Using higher frequencies would imply a higher degree of noise. Moreover, it would make the HDMR analysis unfeasible from a computational standpoint.

ratios are close to 1, and none of them is significant. That is, we can never reject the hypothesis that the forecasts from the headlines are not different from the full body. This fact holds at all the forecasting horizons, and across all the indicators. This result is striking, as the full text contains considerably more abundant information than the headlines alone. However, it looks like this additional content does not add value to forecast economic variables.

2.4.2 HDMR

Section 2.4.1 showed that headline and full text sentiment have the same forecasting accuracy in predicting economic indicators. However, one could argue that our sentiment measure is too simple, and cannot fully capture the information content of the articles. We therefore use HDMR, a fully-fledged machine learning model for text-based forecasting (Kelly et al., 2021).

We repeat the out-of-sample exercise using the HDMR method, and report our results in the left panel of Table II. Once again, we use rolling windows, forecasting horizons of 1, 3 and 12 months, and the out-of-sample period goes from 2015 : 01 to 2021 : 12.

Even with the HDMR forecasting model, all the RMSE ratios are close to 1. Four ratios are statistically significant, corresponding to the PMI and S&P500 targets. In the case of 1-month ahead S&P500 forecasts, the headlines actually perform *better* than the full body, even though this pattern is reversed at the quarterly horizon.

Since our out-of-sample window includes the Covid period, we may worry that this influences our results. We present some robustness checks in the Appendix A. We split the out-of-sample period between pre-Covid (2015 to 2019, Table VI) and post-Covid (2020 to 2021, Table VII). We can see that our conclusions are not materially affected.

Moreover, one could argue that during periods of higher uncertainty, the full text becomes more relevant than the bare headlines. We proxy for uncertainty by using the VIX. We specialized our analysis, computing the R^2_{OS} statistics only for the top VIX

Table II
RMSE Ratios

This table reports the out-of-sample RMSE ratios for HDMR- and sentiment-based forecasts of macroeconomic indicators. Each forecast is built from $y_{t+\tau} = \boldsymbol{\beta}'\boldsymbol{x}_t + \gamma \boldsymbol{z}_t^s + \epsilon_t$. $y_{t+\tau}$ is a macro-financial indicator, τ months ahead. $\boldsymbol{x}_t$ contains a constant and the top 5 principal components from a large panel of macro-financial series (Stock and Watson, 2012). $\boldsymbol{z}_t^s$ contains either sentiment (first three columns), or the HDMR sufficient statistics (last three columns) from the headlines ($s = h$) or the full body ($s = f$) of *Wall Street Journal* news articles. Each ratio is built as $RMSE_\tau^h / RMSE_\tau^f$. p values, shown in parentheses, are based on Diebold and Mariano (2002) equal predictive accuracy tests, with the Harvey et al. (1997) correction. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively. The out-of-sample period goes from 2015 : 01 to 2021 : 12.

	HDMR			Sentiment		
τ (months)	1	3	12	1	3	12
IP: total	1.00	0.99	0.78	0.94	1.06	1.02
	(0.78)	(0.36)	(0.36)	(0.32)	(0.33)	(0.34)
Emp: total	1.00	1.30	0.99	0.93	1.09	1.01
	(0.92)	(0.32)	(0.42)	(0.32)	(0.33)	(0.33)
U: all	1.00	1.00	1.00	0.93	1.09	1.01
	(0.81)	(0.98)	(0.80)	(0.31)	(0.33)	(0.33)
HStarts: total	0.95	1.01	0.98	0.97	0.92	1.03
	(0.19)	(0.78)	(0.67)	(0.32)	(0.29)	(0.41)
PMI	1.09*	1.00	1.05***	1.05	0.98	0.93
	(0.06)	(0.95)	(0.01)	(0.58)	(0.26)	(0.30)
CPI-ALL	0.95	1.01	0.96	1.01	0.98	0.99
	(0.14)	(0.58)	(0.59)	(0.87)	(0.22)	(0.68)
Real AHE: goods	0.98	1.35	1.08	0.93	1.11	1.02
	(0.18)	(0.31)	(0.39)	(0.30)	(0.33)	(0.32)
FedFunds	1.14	0.94	0.98	1.02	1.01	1.04
	(0.19)	(0.19)	(0.66)	(0.41)	(0.32)	(0.40)
M1	1.00	1.00	0.99	1.01	1.01	1.02
	(0.80)	(0.96)	(0.36)	(0.33)	(0.34)	(0.35)
Ex rate: avg	0.98	1.01	1.06	0.87	1.05	1.00
	(0.72)	(0.93)	(0.40)	(0.29)	(0.33)	(0.84)
SP 500	0.85**	1.06*	0.99	0.98	0.96	0.98
	(0.04)	(0.08)	(0.86)	(0.49)	(0.33)	(0.36)
Consumer expect	1.02	1.04	1.04	1.02	1.12	0.97
	(0.78)	(0.68)	(0.43)	(0.62)	(0.38)	(0.16)

quintiles. We do not find evidence for such a mechanism in the data, as the results described before are virtually unaffected.

2.4.3 Front Page

In Section 2.4.1 and 2.4.2 we built our text-based predictors using the whole set of WSJ articles, including approximately 3,000 articles per month. One could argue that the forecasting power of the headlines does not come from their content, but their number. To tackle this issue we restrict our sample to the front page articles only, with on average 166 documents per month. We estimate our sentiment and HDMR statistics using this smaller dataset, only choosing headlines or full text at a time. We repeat the same out-of-sample analysis and report our results in Table III.

Once again, all the RMSE ratios are close to 1, and only a handful of them are significant. Interestingly, all the significant RMSE ratios are below 1: the front page headlines have *higher* forecasting accuracy. Our results are particularly strong in the case of housing starts and the S&P500 index, over multiple horizons.

Lastly, we compare the first page and the full set of WSJ articles. We build the forecasts first from the headlines of the front page articles, and later from the whole WSJ. Table IV presents the results. Most of the RMSE ratios are close to 1. Restricting the text data to the front page does not dramatically hinder the forecasting accuracy.

2.4.4 Benchmark

So far, we compared the forecasting accuracy of headlines against the full text of news articles. We haven't evaluated the overall ability of text to forecast macroeconomic indicators. Kelly et al. (2021) already show that their HDMR model significantly improves the forecasting performance of the 5 principal components of Stock and Watson (2012). We reproduce their empirical analysis, comparing the RMSE of the following predictive

Table III
RMSE Ratios - Front Page

This table reports the out-of-sample RMSE ratios for HDMR- and sentiment-based forecasts of macroeconomic indicators. These forecasts are built from the text on the first page of the *Wall Street Journal* articles. Each forecast is built from $y_{t+\tau} = \boldsymbol{\beta}'\boldsymbol{x}_t + \gamma \boldsymbol{z}_t^s + \epsilon_t$. $y_{t+\tau}$ is a macro-financial indicator, τ months ahead. $\boldsymbol{x}_t$ contains a constant and the top 5 principal components from a large panel of macro-financial series (Stock and Watson, 2012). $\boldsymbol{z}_t^s$ contains either sentiment (first three columns), or the HDMR sufficient statistics (last three columns) from the headlines ($s = h$) or the full body ($s = f$) of *Wall Street Journal* first page news articles. Each ratio is built as $RMSE_\tau^h / RMSE_\tau^f$. p values, shown in parentheses, are based on Diebold and Mariano (2002) equal predictive accuracy tests, with the Harvey et al. (1997) correction. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively. The out-of-sample period goes from 2015 : 01 to 2021 : 12.

	HDMR			Sentiment		
τ (months)	1	3	12	1	3	12
IP: total	0.97	0.99	0.98	1.03	1.03	0.98
	(0.26)	(0.82)	(0.29)	(0.36)	(0.32)	(0.34)
Emp: total	1.00	1.29	1.01	1.00	0.96	0.93
	(0.96)	(0.32)	(0.32)	(0.47)	(0.33)	(0.33)
U: all	1.00	0.98	0.99**	1.01	0.95	0.94
	(0.66)	(0.37)	(0.03)	(0.34)	(0.33)	(0.33)
HStarts: total	0.87**	0.95***	0.93	1.05	1.00	1.03
	(0.02)	(0.01)	(0.34)	(0.33)	(0.92)	(0.35)
PMI	1.43	1.17	1.00	1.05	1.06	1.37
	(0.00)	(0.11)	(0.96)	(0.20)	(0.31)	(0.32)
CPI-ALL	0.89***	1.03	0.96	1.00	1.21	0.78
	(0.01)	(0.40)	(0.54)	(0.97)	(0.32)	(0.34)
Real AHE: goods	0.98	1.37	1.04	0.97	0.91	0.81
	(0.22)	(0.33)	(0.56)	(0.32)	(0.31)	(0.33)
FedFunds	1.09	1.17	1.04	1.00	1.17	0.87
	(0.45)	(0.33)	(0.22)	(0.62)	(0.32)	(0.34)
M1	1.01	1.01	0.87	1.00*	1.02	0.98
	(0.11)	(0.85)	(0.35)	(0.08)	(0.31)	(0.45)
Ex rate: avg	0.92	0.89**	0.92*	1.00	0.80	0.95
	(0.14)	(0.01)	(0.06)	(0.60)	(0.33)	(0.35)
SP 500	0.82***	0.87	0.96***	1.05	1.00	0.92
	(0.01)	(0.35)	(0.01)	(0.30)	(0.79)	(0.33)
Consumer expect	1.53	1.16	1.18	0.98	0.82	0.95***
	(0.00)	(0.28)	(0.12)	(0.72)	(0.37)	(0.00)

Table IV

RMSE Ratios - Front Page vs All Articles

This table reports the out-of-sample RMSE ratios for HDMR- and sentiment-based forecasts of macroeconomic indicators. These forecasts are built from the text on either the first page of the *Wall Street Journal* or the full set of articles. Each forecast is built from $y_{t+\tau} = \boldsymbol{\beta}'\boldsymbol{x}_t + \gamma \boldsymbol{z}_t^s + \epsilon_t$. $y_{t+\tau}$ is a macro-financial indicator, τ months ahead. $\boldsymbol{x}_t$ contains a constant and the top 5 principal components from a large panel of macro-financial series (Stock and Watson, 2012). $\boldsymbol{z}_t^s$ contains either sentiment (first three columns), or the HDMR sufficient statistics (last three columns) from the headlines of the front page ($s = front$) or the full set of articles ($s = all$) of the *Wall Street Journal*. Each ratio is built as $RMSE_\tau^{front}/RMSE_\tau^{all}$. p values, shown in parentheses, are based on Diebold and Mariano (2002) equal predictive accuracy tests, with the Harvey et al. (1997) correction. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively. The out-of-sample period goes from 2015 : 01 to 2021 : 12.

	HDMR			Sentiment		
τ (months)	1	3	12	1	3	12
IP: total	0.96	1.01	1.25	1.09	1.05	0.91
	(0.11)	(0.77)	(0.35)	(0.33)	(0.32)	(0.33)
Emp: total	1.00	0.99	1.02	1.07	1.02	0.88
	(0.98)	(0.27)	(0.36)	(0.32)	(0.35)	(0.33)
U: all	1.00	0.98	0.99	1.08	1.01	0.89
	(0.56)	(0.15)	(0.25)	(0.32)	(0.40)	(0.33)
HStarts: total	0.92***	0.95**	0.95	1.07	1.04	1.05
	(0.01)	(0.03)	(0.22)	(0.33)	(0.29)	(0.34)
PMI	1.31	1.18*	0.95	0.99	1.03	1.25
	(0.00)	(0.06)	(0.51)	(0.88)	(0.31)	(0.34)
CPI-ALL	0.94	1.02	1.01	0.90	1.25	0.76
	(0.11)	(0.59)	(0.91)	(0.43)	(0.30)	(0.31)
Real AHE: goods	1.00	1.01	0.96	1.09	0.99	0.78
	(0.74)	(0.62)	(0.14)	(0.31)	(0.23)	(0.33)
FedFunds	0.95	1.24	1.06	0.96	1.14	0.84
	(0.37)	(0.29)	(0.43)	(0.33)	(0.32)	(0.33)
M1	1.01*	1.01*	0.88	0.99	0.97	0.97
	(0.06)	(0.09)	(0.35)	(0.33)	(0.29)	(0.32)
Ex rate: avg	0.95	0.88	0.87***	1.19	0.83	0.91
	(0.41)	(0.14)	(0.00)	(0.28)	(0.30)	(0.32)
SP 500	0.96	0.82	0.97	0.93	1.07	0.92
	(0.41)	(0.19)	(0.72)	(0.34)	(0.32)	(0.32)
Consumer expect	1.51	1.12	1.13	0.96	0.88	0.95
	(0.00)	(0.45)	(0.31)	(0.35)	(0.31)	(0.27)

models

$$y_{t+\tau} = \boldsymbol{\beta}'\boldsymbol{x}_t + \gamma \boldsymbol{z}_t^f + \epsilon_t, \tag{2.2}$$

$$y_{t+\tau} = \boldsymbol{\beta}'\boldsymbol{x}_t + \epsilon_t, \tag{2.3}$$

where $\mathbf{z}_t^f$ contains either sentiment or the HDMR sufficient statistics extracted from the full body of news articles. We later compute the RMSE ratios for each macroeconomic series at each horizon.

We present our results in Table V. RMSE ratios smaller than 1 indicate that text-enhanced forecasts have smaller errors than the PC benchmark. Most of the ratios are smaller than 1, especially for the HDMR model, in line with the evidence of Kelly et al. (2021) . However, we fail to reject the null of no difference in forecasting accuracy for most of the ratios, even though interestingly this is not true for consumer expectations. This fact, at odds with the evidence of Kelly et al. (2021), might be due to the different out-of-sample period we use, due to data availability. The larger degree of noise in our estimates might also be related to the Covid period. Untabulated results show that limiting the out-of-sample window to $2015 - 2019$ delivers more RMSE ratios to be statistically different from 1.

2.5 The Model

We interpret our findings through an accuracy-clarity trade-off, in the spirit of Myatt and Wallace (2012). A news producer sends two signals, the headline and the full text, for an underlying economic variable. We assume the former is less accurate, i.e. it has more noise than the latter. Readers are interested in forecasting an economic variable and receive a noisy version of the two signals. The headline signal is clearer, i.e. it displays less noise on the receiver side.

Table V
RMSE Ratios

This table reports the out-of-sample RMSE ratios for HDMR- and sentiment-based forecasts of macroeconomic indicators. Each forecast is built from either $y_{t+\tau} = \boldsymbol{\beta}'\boldsymbol{x}_t + \gamma \boldsymbol{z}_t^f + \epsilon_t$ or $y_{t+\tau} = \boldsymbol{\beta}'\boldsymbol{x}_t + \epsilon_t$. $y_{t+\tau}$ is a macro-financial indicator, τ months ahead. $\boldsymbol{x}_t$ contains a constant and the top 5 principal components from a large panel of macro-financial series (Stock and Watson, 2012). $\boldsymbol{z}_t^f$ contains either sentiment (first three columns), or the HDMR sufficient statistics (last three columns) from the full body of *Wall Street Journal* news articles. Each ratio is built as $RMSE_{\tau}^{f}/RMSE_{\tau}$, where $RMSE$ is the model without text data. p values, shown in parentheses, are based on Diebold and Mariano (2002) equal predictive accuracy tests, with the Harvey et al. (1997) correction. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively. The out-of-sample period goes from 2015 : 01 to 2021 : 12.

	HDMR			Sentiment		
h (months)	1	3	12	1	3	12
IP: total	0.48	0.47	0.67	0.99	0.92	1.00
	(0.32)	(0.31)	(0.28)	(0.36)	(0.33)	(0.56)
Emp: total	0.38	0.56	0.46	1.00	0.87	1.00
	(0.32)	(0.32)	(0.27)	(0.31)	(0.34)	(0.61)
U: all	0.40	0.75	0.42	0.99	0.86	1.00
	(0.32)	(0.31)	(0.27)	(0.32)	(0.35)	(0.84)
HStarts: total	0.32	0.88*	0.42	1.00	1.12	0.97
	(0.32)	(0.10)	(0.29)	(0.26)	(0.17)	(0.40)
PMI	0.45*	0.31	0.50	1.00	1.03	1.15
	(0.06)	(0.20)	(0.19)	(0.43)	(0.33)	(0.23)
CPI-ALL	1.05	0.88	0.55	1.08	0.98	1.04
	(0.20)	(0.29)	(0.30)	(0.35)	(0.35)	(0.15)
Real AHE: goods	0.37	0.73	0.50	0.96	0.84	0.97
	(0.32)	(0.32)	(0.29)	(0.32)	(0.35)	(0.29)
FedFunds	0.97***	0.39	0.41	1.03	0.98	0.97
	(0.01)	(0.30)	(0.23)	(0.60)	(0.33)	(0.32)
M1	1.00**	0.95	1.05	1.00	1.04	1.01**
	(0.05)	(0.21)	(0.58)	(0.90)	(0.42)	(0.04)
Ex rate: avg	0.79	0.47	0.56	0.92	0.93	1.02
	(0.32)	(0.30)	(0.29)	(0.30)	(0.34)	(0.66)
SP 500	1.14	1.30	0.65	1.13	1.07	1.04
	(0.28)	(0.11)	(0.26)	(0.40)	(0.25)	(0.27)
Consumer expect	0.45***	0.51**	0.75***	0.91	0.80	1.05
	(0.00)	(0.03)	(0.00)	(0.25)	(0.24)	(0.91)

Our framework combines elements from the rational inattention (Sims, 2003) and beauty contest (Keynes, 1936) literature, as explored in Hellwig and Veldkamp (2009), Myatt and Wallace (2012) and Kim (2022). We consider a simplified model, which lets us obtain analytical derivations. The readers need to forecast an indicator, and have access to two signals with different variances. There is a cost for paying attention. Moreover, readers are incentivized to focus on what others are reading, as deviating from the general consensus involves a penalty (Morris and Shin, 2002). If enough agents have a high attention cost, the model has an equilibrium where every agent only uses the signal from the headlines.

Our framework is particularly relevant to study the impact on indicators such as the stock market, in which the general consensus affects the future value of the indicator (Allen et al., 2006). We can also use it to study the forecasting problem of an econometrician, who needs to extract signal from text to predict an indicator.

2.5.1 Beauty Contest

Our model is close to Myatt and Wallace (2012). There is a continuum of agents indexed by $l \in [0, 1]$, who are interested in forecasting a variable y (e.g. the stock market return). A news-producer sends two signals

$$\bar{y}_s = y + \eta_s, \quad \eta_s \sim N\left(0, k_s^2\right), \tag{2.4}$$

where $s = h$ stands for "headlines" and $s = f$ for "full text". We refer to $1/k_s^2$, the precision of signal s, as its accuracy (Myatt and Wallace, 2012) or content (Banerjee et al., 2022). We assume that the full text is more accurate than the headline, $k_f^2 < k_h^2$, even though it is not necessary for our results.

In line with Myatt and Wallace (2012), each reader l receives a noisy version of the news signals, and decides how much attention to allocate to them. In particular, he receives

$$y_{s,l} = \bar{y}_s + \epsilon_{s,l}, \quad \epsilon_{s,l} \sim N\left(0, \frac{\xi_s^2}{z_{s,l}}\right). \tag{2.5}$$

The precision of $y_{s,l}$ depends on two terms. $z_{s,l}$ is the attention that reader l pays to signal s: ceteris paribus, a larger degree of attention leads to a more precise signal. We refer to $1/\xi_s^2$ as the clarity of source s. We assume that $\xi_f^2 > \xi_h^2$, so that the headline is clearer than the full text.

To simplify our argument we make two assumptions. First, $\xi_h^2 = 0$, which means that the headlines are costless to receive and, as a consequence, every agent reads them. In this model headlines are a public signal in the spirit of Morris and Shin (2002), which is equivalent to having a prior distribution $y \sim N\left(\bar{y}_h, k_h^2\right)$. We also assume that $z_{f,l} \in \{0, 1\}$. If $z_{f,l} = 0$, agent l pays no attention to the full text, and the signal he receives is only noise. As a consequence, he forecasts y only using the headline $\bar{y}_h$. We model attention only at the extensive margin, studying whether or not a reader uses the full text or only uses the headline. This simplified model results in a unique equilibrium. Modeling a continuous level of attention leads to less informative results, as we can find multiple equilibria.

The agents play a simultaneous move game. First, they decide how much attention to devote to each information source. Second, they observe the signals given their attention allocation. Third, they provide their best forecast of indicator y, which we denote a_l in line with the literature. As already mentioned, we model a beauty contest environment, in which the payoffs to each agent depend not only on their proximity to the true value, but also to the average actions of the others. That is, the utility of reader l is

$$u_l = \bar{u} - (1-\gamma)\left(a_l - y\right)^2 - \gamma\left(a_l - \bar{a}\right)^2 - 2cI\left(y_l, \bar{y}_l\right). \tag{2.6}$$

γ is the beauty contest parameter, and determines the incentive . We assume $0 \leq \gamma < 1$, i.e. there are strategic complementarities, the agents want to do what the others are doing. $\bar{a} := \int_0^1 a_l dl$ is the average action across players. Notice that, since agent l has mass 0, it cannot influence it.

To solve the model, we only need to look for linear strategies in the form $A_l(\boldsymbol{y}_l) = w_{h,l}y_h + w_{f,l}y_{f,l}$, with the constraint that $w_{h,l} + w_{f,l} = 1$. That is, the strategy is a weighted average of the signals received by the player. We explain the rationale for this choice in Appendix B.1. In such case the expected utility is

$$\min_{z,w} \sum_{s=h,f} w_{s,l}^2 \left((1-\gamma) k_s^2 + \frac{\xi_s^2}{z_{s,l}} \right) + \gamma \sum_{s=h,f} (w_{s,l} - w_s)^2 k_s^2 + c \log \left(1 + \frac{\left(k_f^2 + k_h^2\right)}{\xi_f^2 / z_{f,l}} \right). \quad (2.7)$$

Notice that the second term can be neglected while minimizing, as any deviation from a symmetric equilibrium is strongly penalized by mean of that beauty contest element. That is, we will look for a symmetric equilibrium. See Appendix B.2 for more details.

We are considering a discrete attention variable, that lets us focus on the intensive margin and makes the whole argument easier. We just need to focus on two cases. In the case of $z_f = 0$, we have $w_f = 0$, that is the full body is not used and, as a consequence, $w_h = 1$. The expected utility in this case is

$$(1-\gamma) k_h^2. \quad (2.8)$$

Consider instead the case where $z_f = 1$, i.e. the reader decides to devote his attention to the full article. It is possible to derive the expressions for the full weights, which give rise to the following Proposition, for which we present details in Appendix B.3.

Proposition 4 (Equilibrium)**.** *There is a unique symmetric equilibrium. There exists* $\bar{c} \geq 0$ *such that* $z_f = 0$ *iff* $c > \bar{c}$. *Such threshold cost is decreasing in* γ.

2.5.2 The Econometric Framework

We consider now the case of an econometrician who receives two signals, and uses them to forecast an indicator. We want to forecast $y \sim N(\mu, \sigma^2)$. In the spirit of our accuracy/clarity tradeoff, the econometrician has two signals: the headline y_h, and the full text y_f. The full text signal he receives is

$$y_f = \bar{y}_f + \epsilon_f, \quad \epsilon_f \sim N\left(0, \xi_f^2\right), \tag{2.9}$$

$$\bar{y}_f = y + \eta_f, \quad \eta_f \sim N\left(0, k_f^2\right), \tag{2.10}$$

while the headline signal is

$$y_h = \bar{y}_h = y + \eta_h, \quad \eta_h \sim N\left(0, k_h^2\right). \tag{2.11}$$

The signals $\bar{y}_h, \bar{y}_f$ are sent by a news producer. The econometrician needs to analyze them, and the signal he extracts is y_h and y_f. Even though we assume that $k_f^2 < k_h^2$, i.e. the full text is more accurate, it has an additional noise component. It is harder to extract a signal from the full text.

We can show that, conditional on receiving a signal, the expected loss is

$$\mathbb{E}\left[(y - \hat{y}_s)^2\right] = \frac{\sigma^2 \sigma_s^2}{\sigma^2 + \sigma_s^2}, \tag{2.12}$$

$\sigma_f^2 = k_f^2 + \xi_f^2$ and $\sigma_h^2 = k_h^2$. We can show that

Proposition 5 (Forecaster Problem)**.** *The expected loss from the full text is smaller than the one for the headlines iff*

$$\frac{1}{\sigma_f^2} > \frac{1}{k_h^2 - k_f^2}, \tag{2.13}$$

that is if the clarity of the full text is large enough. The larger the accuracy of the headlines, $1/k_h^2$, the harder it is to switch to the full text.

2.6 Conclusion

In this chapter, we compare the information contained in the headlines and the full text of *Wall Street Journal* articles. Our findings reveal a strong correlation between the sentiment measures extracted from the two sources. We conduct an out-of-sample analysis, forecasting macro-financial indicators using either sentiment or a state-of-the-art machine learning model. We find that the predictors based on the headlines and full text have a similar forecasting accuracy. If anything, it appears that the headlines lead to more accurate forecasts. We interpret our findings in a model that includes both costs of attention and beauty contest elements.

A Robustness

Table VI
RMSE Ratios - Pre Covid

This table reports the out-of-sample RMSE ratios for HDMR- and sentiment-based forecasts of macroeconomic indicators. Each forecast is built from $y_{t+\tau} = \boldsymbol{\beta}'\boldsymbol{x}_t + \gamma \boldsymbol{z}_t^s + \epsilon_t$. $y_{t+\tau}$ is a macro-financial indicator, τ months ahead. $\boldsymbol{x}_t$ contains a constant and the top 5 principal components from a large panel of macro-financial series (Stock and Watson, 2012). $\boldsymbol{z}_t^s$ contains either sentiment (first three columns), or the HDMR sufficient statistics (last three columns) from the headlines ($s = h$) or the full body ($s = f$) of *Wall Street Journal* news articles. Each ratio is built as $RMSE_\tau^h / RMSE_\tau^f$. p values, shown in parentheses, are based on Diebold and Mariano (2002) equal predictive accuracy tests, with the Harvey et al. (1997) correction. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively. The out-of-sample period goes from 2015 : 01 to 2019 : 12.

	HDMR			Sentiment		
τ (months)	1	3	12	1	3	12
IP: total	1.12*	1.12	0.99	1.01	1.00	1.02
	(0.08)	(0.10)	(0.82)	(0.59)	(0.82)	(0.40)
Emp: total	0.90	1.06	1.03	1.03	1.00	1.01
	(0.12)	(0.13)	(0.80)	(0.20)	(0.89)	(0.43)
U: all	0.98	0.98	1.04	0.99	1.00	1.01
	(0.78)	(0.79)	(0.78)	(0.68)	(0.75)	(0.70)
HStarts: total	0.90*	0.99	0.96	0.99	1.00	0.96*
	(0.08)	(0.88)	(0.67)	(0.62)	(0.43)	(0.09)
PMI	1.04	0.97	1.08	0.99	0.99	0.99
	(0.33)	(0.66)	(0.10)	(0.58)	(0.84)	(0.69)
CPI-ALL	0.90*	1.00	1.05**	0.97*	0.99	1.05
	(0.05)	(0.93)	(0.04)	(0.07)	(0.63)	(0.28)
Real AHE: goods	1.00	1.03	1.10	0.98	0.98	1.01
	(0.96)	(0.77)	(0.73)	(0.22)	(0.23)	(0.20)
FedFunds	0.94*	1.05**	1.07***	0.99	1.00	0.98
	(0.06)	(0.03)	(0.00)	(0.64)	(0.81)	(0.70)
M1	0.94	1.00	1.00	1.00	1.00	1.01
	(0.38)	(1.00)	(0.98)	(0.89)	(0.85)	(0.90)
Ex rate: avg	0.97	1.07	1.10	1.00	0.99	1.03
	(0.68)	(0.44)	(0.39)	(0.97)	(0.45)	(0.60)
SP 500	0.96	1.06	1.11***	1.00	1.00	1.01
	(0.45)	(0.33)	(0.00)	(0.99)	(0.79)	(0.45)
Consumer expect	0.97	0.88	0.98	1.01	0.99	0.96
	(0.64)	(0.14)	(0.77)	(0.75)	(0.58)	(0.31)

Table VII

RMSE Ratios - Post Covid

This table reports the out-of-sample RMSE ratios for HDMR- and sentiment-based forecasts of macroeconomic indicators. Each forecast is built from $y_{t+\tau} = \boldsymbol{\beta}'\boldsymbol{x}_t + \gamma \boldsymbol{z}_t^s + \epsilon_t$. $y_{t+\tau}$ is a macro-financial indicator, τ months ahead. $\boldsymbol{x}_t$ contains a constant and the top 5 principal components from a large panel of macro-financial series (Stock and Watson, 2012). $\boldsymbol{z}_t^s$ contains either sentiment (first three columns), or the HDMR sufficient statistics (last three columns) from the headlines ($s = h$) or the full body ($s = f$) of *Wall Street Journal* news articles. Each ratio is built as $RMSE_\tau^h / RMSE_\tau^f$. p values, shown in parentheses, are based on Diebold and Mariano (2002) equal predictive accuracy tests, with the Harvey et al. (1997) correction. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively. The out-of-sample period goes from 2020 : 01 to 2021 : 12.

	HDMR			Sentiment		
τ (months)	1	3	12	1	3	12
IP: total	1.00	0.98	0.77	0.94	1.06	1.02
	(0.99)	(0.25)	(0.22)	(0.33)	(0.34)	(0.24)
Emp: total	1.00	1.30	0.99	0.93	1.09	1.01
	(0.91)	(0.32)	(0.57)	(0.32)	(0.33)	(0.13)
U: all	1.00	1.00	1.00	0.93	1.09	1.01
	(0.80)	(0.98)	(0.88)	(0.32)	(0.33)	(0.14)
HStarts: total	0.98	1.02	1.00	0.96	0.90	1.03*
	(0.79)	(0.40)	(0.97)	(0.33)	(0.30)	(0.05)
PMI	1.15	1.02	1.03	1.07	0.98	0.92
	(0.12)	(0.75)	(0.25)	(0.57)	(0.25)	(0.17)
CPI-ALL	1.03	1.03	0.85	1.02	0.96	0.98
	(0.49)	(0.68)	(0.19)	(0.65)	(0.24)	(0.38)
Real AHE: goods	0.98	1.37	1.08	0.93	1.11	1.02*
	(0.18)	(0.31)	(0.26)	(0.31)	(0.33)	(0.05)
FedFunds	1.20	0.92*	0.95	1.03	1.01	1.04
	(0.16)	(0.07)	(0.37)	(0.39)	(0.33)	(0.31)
M1	1.00	1.00	0.99	1.01	1.01	1.02
	(0.86)	(0.97)	(0.89)	(0.34)	(0.34)	(0.31)
Ex rate: avg	1.01	0.84	0.99	0.73	1.06	1.00
	(0.97)	(0.12)	(0.85)	(0.30)	(0.31)	(0.15)
SP 500	0.79*	1.07	0.91	0.98	0.93	0.97**
	(0.06)	(0.12)	(0.12)	(0.48)	(0.31)	(0.04)
Consumer expect	1.04	1.13	1.08**	1.03	1.23	0.99
	(0.57)	(0.32)	(0.04)	(0.67)	(0.37)	(0.84)

Table VIII
RMSE Ratios - Lead Paragraph vs Full Article

This table reports the out-of-sample RMSE ratios for HDMR- and sentiment-based forecasts of macroeconomic indicators. We replace headlines with the lead paragraph of each article. Each forecast is built from $y_{t+\tau} = \boldsymbol{\beta}'\boldsymbol{x}_t + \gamma \boldsymbol{z}_t^s + \epsilon_t$. $y_{t+\tau}$ is a macro-financial indicator, τ months ahead. $\boldsymbol{x}_t$ contains a constant and the top 5 principal components from a large panel of macro-financial series (Stock and Watson, 2012). $\boldsymbol{z}_t^s$ contains either sentiment (first three columns), or the HDMR sufficient statistics (last three columns) from the headlines ($s = h$) or the full body ($s = f$) of *Wall Street Journal* news articles. Each ratio is built as $RMSE_\tau^h / RMSE_\tau^f$. p values, shown in parentheses, are based on Diebold and Mariano (2002) equal predictive accuracy tests, with the Harvey et al. (1997) correction. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively. The out-of-sample period goes from 2020 : 01 to 2021 : 12.

	HDMR			Sentiment		
h (months)	1	3	12	1	3	12
IP: total	1.03	0.99	0.98	0.94	1.06	1.00
	(0.32)	(0.31)	(0.23)	(0.33)	(0.33)	(0.33)
Emp: total	1.00	1.33	1.00	0.94	1.07	0.99
	(0.64)	(0.31)	(0.91)	(0.32)	(0.33)	(0.33)
U: all	1.00	1.01	1.00	0.93	1.08	0.99
	(0.86)	(0.30)	(0.90)	(0.32)	(0.33)	(0.33)
HStarts: total	1.00	0.96	0.99	0.97	0.95	1.02
	(0.97)	(0.29)	(0.94)	(0.33)	(0.29)	(0.34)
PMI	1.04	0.95	0.94	1.05	1.01	1.00
	(0.43)	(0.17)	(0.16)	(0.59)	(0.43)	(0.93)
CPI-ALL	0.95	1.11	0.93	0.99	0.99**	1.02
	(0.17)	(0.27)	(0.21)	(0.62)	(0.04)	(0.17)
Real AHE: goods	0.99	1.39	1.05	0.93	1.09	0.99
	(0.72)	(0.30)	(0.43)	(0.30)	(0.32)	(0.31)
FedFunds	1.00	0.99	1.02	1.02	1.01	1.00
	(0.98)	(0.47)	(0.35)	(0.31)	(0.30)	(0.36)
M1	1.00	0.98	1.28	1.01	1.02	1.03
	(0.34)	(0.58)	(0.33)	(0.33)	(0.32)	(0.32)
Ex rate: avg	0.95	0.99	0.87	0.82	1.05	1.02
	(0.42)	(0.91)	(0.18)	(0.28)	(0.33)	(0.19)
SP 500	0.86*	1.03	1.01	0.99	0.98	1.03
	(0.07)	(0.36)	(0.88)	(0.70)	(0.53)	(0.29)
Consumer expect	1.14*	1.08	1.11	1.04	1.08	1.00
	(0.08)	(0.29)	(0.12)	(0.31)	(0.32)	(0.93)

Table IX

RMSE Ratios - Headlines vs Lead Paragraphs

This table reports the out-of-sample RMSE ratios for HDMR- and sentiment-based forecasts of macroeconomic indicators. We replace headlines with the lead paragraph of each article. Each forecast is built from $y_{t+\tau} = \boldsymbol{\beta}'\boldsymbol{x}_t + \gamma \boldsymbol{z}_t^s + \epsilon_t$. $y_{t+\tau}$ is a macro-financial indicator, τ months ahead. $\boldsymbol{x}_t$ contains a constant and the top 5 principal components from a large panel of macro-financial series (Stock and Watson, 2012). $\boldsymbol{z}_t^s$ contains either sentiment (first three columns), or the HDMR sufficient statistics (last three columns) from the headlines ($s = h$) or the full body ($s = f$) of *Wall Street Journal* news articles. Each ratio is built as $RMSE_\tau^h/RMSE_\tau^f$. p values, shown in parentheses, are based on Diebold and Mariano (2002) equal predictive accuracy tests, with the Harvey et al. (1997) correction. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively. The out-of-sample period goes from 2020 : 01 to 2021 : 12.

	HDMR			Sentiment		
h (months)	1	3	12	1	3	12
IP: total	0.98	0.99	0.79	1.00	1.00	1.00
	(0.51)	(0.70)	(0.34)	(0.48)	(0.34)	(0.99)
Emp: total	1.00	0.98	0.99	1.00	1.00	1.00
	(0.72)	(0.24)	(0.40)	(0.75)	(0.32)	(0.37)
U: all	1.00	0.99	1.00	1.00	1.00	1.00
	(0.99)	(0.22)	(0.86)	(0.91)	(0.23)	(0.35)
HStarts: total	0.94	1.05	0.99	1.00**	1.00	1.00
	(0.18)	(0.18)	(0.86)	(0.05)	(0.61)	(0.38)
PMI	1.04	1.04	1.12***	1.00	1.00	1.00
	(0.48)	(0.31)	(0.00)	(0.80)	(0.32)	(0.38)
CPI-ALL	1.00	0.91	1.03	1.00	1.00	1.00
	(0.95)	(0.29)	(0.84)	(0.33)	(0.79)	(0.59)
Real AHE: goods	0.99	0.97	1.03	1.00	1.00	1.00
	(0.50)	(0.30)	(0.22)	(0.26)	(0.28)	(0.29)
FedFunds	1.14	0.96	0.96	1.00	1.00	1.00
	(0.15)	(0.28)	(0.37)	(0.36)	(0.40)	(0.63)
M1	1.00	1.02	0.77	1.00	1.00	1.00
	(0.73)	(0.58)	(0.33)	(0.71)	(0.32)	(0.36)
Ex rate: avg	1.02	1.01	1.21**	1.00	1.00	1.00
	(0.75)	(0.87)	(0.02)	(0.21)	(0.31)	(0.15)
SP 500	0.99	1.03	0.98	1.00	1.00	1.00
	(0.78)	(0.59)	(0.60)	(0.72)	(0.10)	(0.97)
Consumer expect	0.89*	0.96	0.94	1.00	1.00	1.00
	(0.09)	(0.56)	(0.45)	(0.55)	(0.25)	(0.92)

B The Model

Throughout the model, the strategy that a player chooses is given by $\{z_l, A(y_l)\}$. We omit the vector notation for the sake of clarity.

B.1 Linearity of Strategies

In the main text we claimed that $A(y_l)$ is linear in the signal y_l, in line with the literature. We now motivate this claim. Suppose that any other player l' follows a strategy in the form $\{z_{l'}, A(y_{l'})\}$. Taking the expected utility of player l, and computing the first order conditions, we get

$$\frac{\partial \mathbb{E}\left[u_l \mid y_l\right]}{\partial a_l} = -2\left(1-\gamma\right)\left(a_l - \mathbb{E}\left[y \mid y_l\right]\right) - 2\gamma \mathbb{E}\left[a_l - \bar{a} \mid y_l\right] = 0, \tag{14}$$

which delivers

$$a_l = A\left(y_l\right) = \left(1-\gamma\right)\mathbb{E}\left[y \mid y_l\right] + \gamma \mathbb{E}\left[A\left(y_{l'}\right) \mid y_l\right]. \tag{15}$$

In the last line we used

$$\mathbb{E}\left[\bar{a} \mid y_l\right] = \mathbb{E}\left[A\left(y_{l'}\right) \mid y_l\right], \quad \forall l' \neq l. \tag{16}$$

Indeed, since $\bar{a} = \int dl' a_{l'}$, we have $\mathbb{E}\left[\bar{a} \mid y_l\right] = \int dl' \mathbb{E}\left[a_{l'} \mid y_l\right]$. As $\mathbb{E}\left[a_{l'} \mid y_l\right]$ is independent of dl', we can take it out of the integral. We later use $a_{l'} = A\left(y_{l'}\right)$

y follows a normal distribution, and by writing the posterior distribution $y \mid y_l$ we can see that $\mathbb{E}\left[y \mid y_l\right]$ is linear in y_l. Moreover, since $A()$ is linear, also $\mathbb{E}\left[A\left(y_{l'}\right) \mid y_l\right]$ is linear in y_l. As a consequence, the best reply of player l is still linear in the signal realization.

Having shown that we can use linear strategies, we now prove that the weights sum up to 1. The linearity of A implies that we can write $A\left(y_l\right) = w' y_l$, with the vector of weights $w \in \mathbb{R}^2$, as we have only two signals. Therefore, $\mathbb{E}\left[A\left(y_{l'}\right) \mid y_l\right] = w' \mathbb{E}\left[y_{l'} \mid y_l\right]$. Because of normality, we can write the latter expectation, studying the posterior distribution, as $E\left[y_{l'} \mid y_l\right] = By_l$, with B being a 2×2 matrix whose rows sum to 1. In the same spirit, we have that $\mathbb{E}\left[y \mid y_l\right] = a'y_l$, with the elements of a summing to 1. If we compute the optimal action via Equation (15), we get

$$w'y_l = A\left(y_l\right) = \left(1-\gamma\right)\mathbb{E}\left[y \mid y_l\right] + \gamma\mathbb{E}\left[A\left(y_{l'}\right) \mid y_l\right] \tag{17}$$

$$= \left(1-\gamma\right)a'y_l + \gamma w'By_l \tag{18}$$

$$= \left[\left(1-\gamma\right)a + \gamma B'w\right]'y_l, \tag{19}$$

so that

$$w = \left(1-\gamma\right)a + \gamma B'w. \tag{20}$$

Since the elements of a sum to 1, and the columns of B' sum to 1, Equation (20) implies that the elements of w must sum to 1 as well. To see it more clearly, we can write

$$B = \begin{pmatrix} b_1 & 1-b_1 \\ b_2 & 1-b2 \end{pmatrix} \tag{21}$$

and, carrying out the calculations, we must have $w'\mathbf{1} = 1$ with $\mathbf{1}' = (1,1)$.

B.2 Expected Utility

We gather the shocks that each agent receives in the vector $\eta' = \left(\eta_h, \eta_f\right)$ and $\epsilon_l = \left(\epsilon_{l,h}, \epsilon_{l,f}\right)$. We can thus express a_l as

$$a_l = w_l'y_l = w_l'\left(\bar{y} + \epsilon_l\right) = w_l'\left(y\mathbf{1} + \eta + \epsilon_l\right), \tag{22}$$

in which $\mathbf{1}' = (1,1)$. Since the weights sum to 1, $w_l'\mathbf{1} = 1$, we have

$$a_l - y = w_l'(\eta + \epsilon_l) = \sum_{s\in\{h,f\}} w_{s,l}(\eta_s + \epsilon_{s,l}). \tag{23}$$

The previous equation implies that

$$\mathbb{E}\left[(a_l - y)^2\right] = \mathbb{E}\left[\left(\sum_{s\in\{h,f\}} w_{s,l}(\eta_s + \epsilon_{s,l})\right)^2\right] = \sum_{s\in\{h,f\}} w_{s,l}^2\left(k_s^2 + \frac{\xi_{s,l}^2}{z_{s,l}}\right), \tag{24}$$

as the error terms are uncorrelated with each other.

Because of the law of large numbers, the average action is

$$\bar{a} = \int dl' a_{l'} = \int dl' w_l'(y\mathbf{1} + \eta + \epsilon_{l'}) = y + \sum_{s\in\{h,f\}} w_s\eta_s. \tag{25}$$

Indeed, by the law of large numbers $\int dl' w_l' \epsilon_{l'} = 0$. We thus have

$$a_l - \bar{a} = \sum_{s\in\{h,f\}} w_{s,l}\epsilon_{s,l} + \sum_{s\in\{h,f\}} (w_{s,l} - w_s)\eta_s, \tag{26}$$

and we get the expected value

$$\mathbb{E}\left[(a_l - \bar{a})^2\right] = \sum_{s\in\{h,f\}} \frac{w_{s,l}^2\xi_{s,l}^2}{z_{s,l}} + \sum_{s\in\{h,f\}} (w_{s,l} - w_s)^2 k_s^2. \tag{27}$$

The expected utility is thus

$$\begin{aligned}\mathbb{E}[u_l] = \bar{u} - \sum_{s=h,f} w_{s,l}^2\left((1-\gamma)k_s^2 + \frac{\xi_s^2}{z_{s,l}}\right) + \\ -\gamma\sum_{s=h,f}(w_{s,l} - w_s)^2 k_s^2 - c\log\left(1 + \frac{\left(k_f^2 + k_h^2\right)}{\xi_f^2/z_{f,l}}\right).\end{aligned} \tag{28}$$

The last term comes from $I(y_l, \bar{y}_l)$, since we have gaussian noise. Indeed, if two variables

X and Y are jointly gaussian, with correlation ρ, we have $I(X,Y) = \frac{1}{2}\log\frac{1}{1-\rho^2}$. We can thus compute

$$Cov\left(y_{f,l}, \bar{y}_f\right) = \mathbb{E}\left[y_{f,l}\bar{y}_f\right] - \mathbb{E}\left[y_{f,l}\right]\mathbb{E}\left[\bar{y}_f\right] = \mathbb{E}\left[y^2 + \eta_f^2\right] - \mathbb{E}\left[y^2\right] = k_h^2 + k_f^2. \tag{29}$$

We also have

$$\sigma^2\left(y_{f,l}\right) = k_h^2 + k_f^2 + \frac{\xi_f^2}{z_{f,l}}, \tag{30}$$

$$\sigma^2\left(\bar{y}_f\right) = k_h^2 + k_f^2, \tag{31}$$

which implies

$$I\left(y_{f,l}, \bar{y}_f\right) = \frac{1}{2}\log\frac{1}{1-\rho^2} = \frac{1}{2}\log\frac{1}{1-\frac{k_h^2+k_f^2}{k_h^2+k_f^2+\xi_f^2/z_{f,l}}} = \frac{1}{2}\log\left(1 + \frac{\left(k_f^2+k_h^2\right)}{\xi_f^2/z_{f,l}}\right). \tag{32}$$

B.3 Equilibrium

To prove that there is a unique symmetric equilibrium, we start by showing that all equilibria in the model are symmetric. Let us recall that the agents solve the minimization problem

$$\min_{z,w}\sum_{s=h,f} w_{s,l}^2\left(\left(1-\gamma\right)k_s^2 + \frac{\xi_s^2}{z_{s,l}}\right) + \gamma\sum_{s=h,f}\left(w_{s,l} - w_s\right)^2 k_s^2 + c\log\left(1 + \frac{\left(k_f^2+k_h^2\right)}{\xi_f^2/z_{f,l}}\right). \tag{33}$$

We can limit our research to symmetric equilibria. Indeed, consider a symmetric equilibrium, such that $w_{s,l} = w_s$ for each player l. Because of the beauty contest term (the second term in the previous equation), each agent has no incentive to deviate from such equilibrium. This term represents the loss for a player that uses a strategy different from the strategies of the other players. We are considering the case $\gamma > 0$, so that players have an incentive to coordinate, that in this simplified setting leads to a symmetric equilibrium.

References

Allen, F., Morris, S., and Shin, H. S. (2006). Beauty contests and iterated expectations in asset markets. *The Review of Financial Studies*, 19(3):719–752.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Banerjee, S., Qu, H., and Zhao, R. (2022). Clarity trumps content: An experiment on information acquisition in beauty contests. *Journal of Economic Behavior & Organization*, 195:381–407.

Bianchi, F., Kung, H., and Cram, R. G. (2021). Using social media to identify the effects of congressional partisanship on asset prices. Technical report, National Bureau of Economic Research.

Bybee, L., Kelly, B. T., Manela, A., and Xiu, D. (2021). Business news and business cycles. Technical report, National Bureau of Economic Research.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

Gabielkov, M., Ramachandran, A., Chaintreau, A., and Legout, A. (2016). Social clicks: What and who gets read on twitter? In *Proceedings of the 2016 ACM SIGMETRICS international conference on measurement and modeling of computer science*, pages 179–192.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*, 133(2):801–870.

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting*, 13(2):281–291.

Hassan, T. A., Hollander, S., Van Lent, L., and Tahoun, A. (2019). Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics*, 134(4):2135–2202.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161.

Hellwig, C. and Veldkamp, L. (2009). Knowing what others know: Coordination motives in information acquisition. *The Review of Economic Studies*, 76(1):223–251.

Huang, D., Jiang, F., Tu, J., and Zhou, G. (2015). Investor sentiment aligned: A powerful predictor of stock returns. *The Review of Financial Studies*, 28(3):791–837.

Kelly, B., Manela, A., and Moreira, A. (2021). Text selection. *Journal of Business & Economic Statistics*, 39(4):859–879.

Keynes, J. M. (1936). The general theory of interest, employment and money.

Kim, J. H. J. (2022). Beauty contest with rationally inattentive agents. *Available at SSRN 2444456*.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.

Manela, A. and Moreira, A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1):137–162.

Morris, S. and Shin, H. S. (2002). Social value of public information. *american economic review*, 92(5):1521–1534.

Myatt, D. P. and Wallace, C. (2012). Endogenous information acquisition in coordination games. *The Review of Economic Studies*, 79(1):340–374.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2020). Measuring news sentiment. *Journal of econometrics*.

Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690.

Stock, J. H. and Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4):481–493.

Chapter 3

Mispricing Proxies

This chapter is based on the paper "Mispricing Proxies", which is joint work with Carlo Ambrogio Favero and Ilaria Leoni.

3.1 Introduction

We investigate whether measures of sentiment extracted from quarterly earnings conference-calls affect the dynamics of stock prices. Using a cross-section of publicly traded companies, we show that sentiment positively correlates with price deviations from their long-run trend, estimated via an error correction model. We document that even though sentiment does not predict future stock returns, it impacts the speed at which prices revert to equilibrium. We find asymmetric effects on overpriced and underpriced stocks.

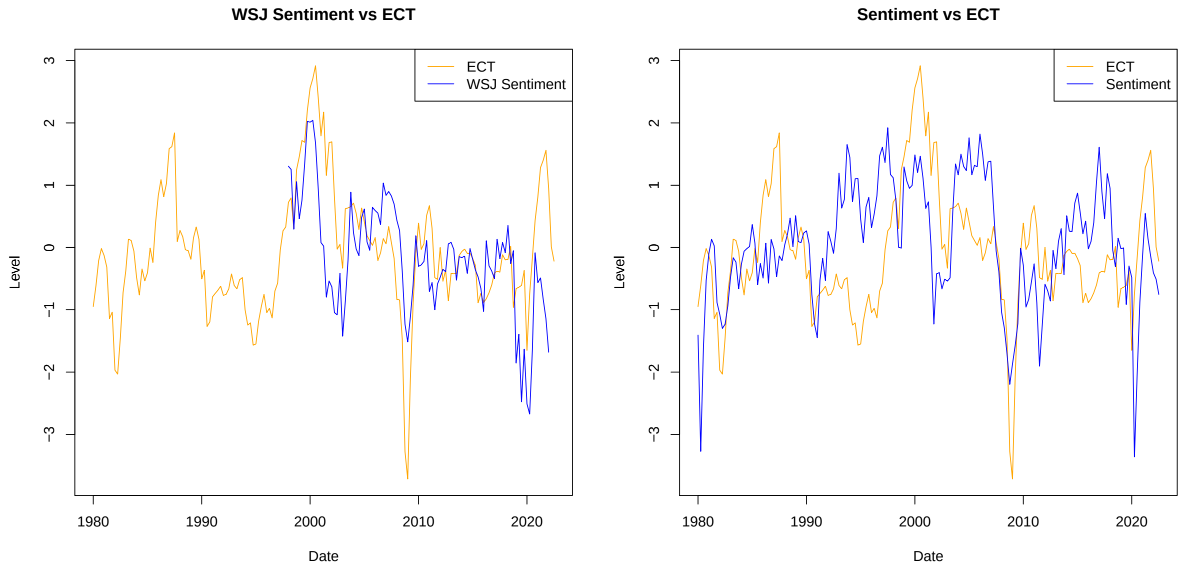
Financial markets are characterized by frequent deviations from their long-run equilibrium (Favero et al., 2019). These deviations, which we interpret as mispricings[1], can result from various factors, such as overreaction to news (Gennaioli and Shleifer, 2018), financial frictions (Duffie, 2010), or changes in market fundamentals. Extending standard factor models to incorporate these deviations improves their forecasting and asset

[1]We use a framework in which deviations from a long-run equilibrium reflect mispricing, as in Dong et al. (2022). For the moment, we do not consider a risk-based environment.

allocation performance. It is natural to investigate the role of investor sentiment in this class of long-run asset pricing models. We explore the relationship between sentiment and price dynamics drawing inspiration from the "diagnostic expectations" literature (Bordalo et al., 2019). Accordingly, we interpret the deviations of prices from their long-run trends as a consequence of investors' overreaction to news. On the one hand, sentiment may shed light on the channels that explain prices deviations from their long-run trends. On the other, estimates of mispricing are often only available at low frequencies. Measures of sentiment, which can be extracted at much higher frequencies, can help overcome this limitation.

We provide some graphical intuition in Figure 3.1, which includes sentiment measures vis-a-vis a measure of aggregate market mispricing. We compute the latter as an error correction term (ECT), i.e. as the residual of the regression of the log-price of the aggregate market on macroeconomic variables (Favero et al., 2019). The left panel (a) in Figure 3.1 shows the sentiment computed from online *Wall Street Journal* articles. This series highly correlates with the ECT, with a correlation coefficient of 0.52. The right panel (b) presents the sentiment indicator computed by Shapiro et al. (2022), available over a longer period. The correlation between this indicator and the ECT is 0.34. Overall, the results depicted in Figure 3.1 suggest that, at least for the aggregate market, sentiment and ECT are correlated.

We carry out our empirical analysis considering a cross-section of publicly traded companies. In particular, we focus on the companies in the Dow Jones Industrial Average (DJIA) index. We construct a granular measure of sentiment for each firm in our sample by analyzing their quarterly earnings conference-calls. Even though the earnings calls are available at a lower frequency than business news, they let us treat each company consistently, as some firms are sistematically overrepresented in the news (e.g. Apple). Earnings calls transcripts are available for each company in our sample, and they are widely followed and covered by financial markets and the economic literature (Hassan



(a) WSJ Sentiment **(b)** Sentiment

Figure 3.1: The picture depicts some mispricing and sentiment measures. Panel (a) uses a measure of sentiment extracted from *Wall Street Journal* headlines, while panel (b) uses the sentiment measure from Shapiro et al. (2022). The error correction term ECT in both panels is obtained as the residual from the regression of market levels on macroeconomic variables.

et al., 2019). We compute the sentiment for each document using a state-of-the-art natural language processing (NLP) model, FinBERT (Araci, 2019).

Our measure of company-level mispricing is obtained through an error correction framework (Favero et al., 2019). We use an econometric model in which prices are driven by a permanent and a temporary component, as in Fama and French (1988). The permanent component, or long-run trend, is common in the cross-section of prices. We estimate it using the values of buy-and-hold portfolios that reproduce the 5 Fama French factors (Fama and French, 2015). The temporary component is asset specific, and we construct it as the residual of a regression of the log-price of a company against the long-run trend. If prices and trends are cointegrated, this residual is stationary and it can be interpreted as an error correction term (ECT), which we will refer to as mispricing.

We compare our series of sentiment and ECTs for the companies in the DJIA. We consider a fixed effect model, and we show that the two variables are positively correlated in the cross-section, in line with what we described for the aggregate market in Figure 3.1. We later compare them in a set of predictive regressions. In particular, we regress the quarterly return of each company against its ECT or sentiment in the previous quarter. We find that most of our estimates of the ECT coefficients are negative and statistically significant. This fact is in line with an error correction framework. Indeed, suppose that the price of an asset in quarter t is above its long-run trend, i.e. the ECT is positive. If the ECT coefficient in the predictive regression is negative, a positive ECT implies a lower return in quarter $t + 1$, so that the price reverts to its long-run trend. On the contrary, we find that the sentiment coefficients in the predictive regressions are not statistically different from zero, and in some cases they are even positive. This fact shows that, even though sentiment and ECT appear to be correlated, we cannot simply use the former as a proxy for the latter.

We later show that sentiment enters the error correction framework through the speed of adjustment. Heuristically, if an asset is overpriced but investor sentiment remains high, it will take longer to revert to equilibrium. We can frame this idea through a model of diagnostic expectations in which investors' beliefs are confirmed by the sentiment signal (Bordalo et al., 2021). We regress the quarterly returns of each company against their lagged ECT, and we impose that the speed of adjustment, i.e. the ECT coefficient, is a linear function of sentiment. We account for an asymmetric reaction to positive and negative mispricings. Our findings show that, when the price is above its equilibrium level, a larger sentiment lowers the speed of adjustment, so that it will take more periods to correct the overpricing shocks. We only find weak effects in the case of negative ECTs. We further investigate the asymmetric effects of sentiment on price dynamics. We show that this asymmetry holds conditionally but not unconditionally. In our sample, a model with constant speed of adjustment would estimate the same coefficient for positive and negative ECTs. On the contrary, allowing for time varying speed of adjustments shows asymmetric effects.

Finally, our main dataset and empirical strategy is at the quarterly frequency, due the release schedule of the earnings calls. However, we potentially have more text information at higher frequency from business news. In Appendix B we analyse whether this information can be useful in the context of forecasting and asset allocation, within a long-run error correction model, using a mixed data sampling (MIDAS) approach.

The remainder of the chapter is organized as follows. Section 3.2 describes the methodology used in our analysis. Section 3.3 presents our data. Section 3.4 presents the empirical results and discusses their implications. Section 3.5 concludes the chapter and provides directions for future research.

3.2 The Empirical Framework

In this section we describe our empirical strategy. We estimate the deviations of prices from their long-run equilibrium via an error correction model, and we compare them to sentiment measures extracted from text data.

3.2.1 The Error Correction Term

We follow Favero et al. (2019) to estimate the deviation of each asset price from its long run equilibrium, which is our measure for mispricing. Let $p_{i,t}$ be the log-price of asset i at time t, at the quarterly frequency. We estimate its deviation from the equilibrium as the residual $u_{i,t}$ from the regression

$$p_{i,t} = \alpha_i t + \boldsymbol{\beta}_i' \log \boldsymbol{F}_t + u_{i,t}. \tag{3.1}$$

Equation (3.1) includes both a deterministic linear trend, t, and a stochastic one, $\log \boldsymbol{F}_t$, which we assume to be shared across assets. In our empirical analysis, we will estimate each component of $\boldsymbol{F}_t$ as the value of a buy-and-hold portfolio that replicates a macro-financial factor. As an example, if we collect the log-returns of the 5 Fama-French factors (Fama and French, 2015) in $\boldsymbol{f}_t$, we have

$$\log \boldsymbol{F}_t - \log \boldsymbol{F}_{t-1} = \boldsymbol{f}_t. \tag{3.2}$$

We will refer to $u_{i,t}$ as the error correction term (ECT). Even though all the variables in Equation (3.1) are non-stationary, they can be shown to be cointegrated, in which case $u_{i,t}$ is stationary (Favero et al., 2019). This cointegration relation lets us express each asset's return between time t and $t+1$ as

$$r_{i,t,t+1} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_{t+1} + \delta_i u_{i,t} + \epsilon_{i,t+1}, \tag{3.3}$$

with $r_{t,t+1} = p_{t+1} - p_t$. If there is cointegration, $\delta_i < 0$ and $u_{i,t}$ has predictive power for the return over the next quarter, which helps improving factor models as in Favero et al. (2019). As an example, suppose that $u_{i,t} < 0$, which means that $p_{i,t}$ is below to its long-run trend. Since $\delta_i < 0$, the ECT term in Equation (3.3) will give a positive contribution, i.e. the next quarter return will be larger. In this way, any deviation from equilibrium is corrected, and we can talk about mispricings.

A special case of the error correction model we described so far applies when the asset we consider is the aggregate market itself. In such case, we will use a different set of factors than the one that describes the cross-section of assets. In particular, let p_t be the log-price of the market ((i.e. the level of an equity index). We will estimate the market deviations from equilibrium via

$$p_t = \boldsymbol{\beta}' \boldsymbol{x}_t + u_t, \tag{3.4}$$

with $\boldsymbol{x}_t = (1, d_t, MY_t, \Delta y_t)'$. d_t is the aggregate dividend, MY_t is the middle-to-young ratio and Δy_t is the potential output growth. The choice of demographic variables to model long-run trends follows Favero et al. (2011). We will later write the predictive equation via

$$r_{t,t+1} = \boldsymbol{\beta}' \boldsymbol{\Delta x}_{t+1} + \delta u_t + \epsilon_{t+1}, \tag{3.5}$$

with $\boldsymbol{\Delta x}_{t+1} = \boldsymbol{x}_{t+1} - \boldsymbol{x}_t$.

3.2.2 Measuring Sentiment

A simple way to estimate sentiment from text data is to use a lexicon approach. It involves using a pre-defined set of words with assigned positive, negative, or neutral sentiment scores. Even though we employed this method to produce the simple plot of Figure 3.1, we rely on a more advanced model to analyze the earnings call transcripts in our main analysis, to properly account for the complexity of financial documents. In particular, we use FinBERT (Araci, 2019). FinBERT (Financial Bidirectional Encoder Representations from Transformers) is a pre-trained Natural Language Processing (NLP) model. It is based on another language model called BERT (Bidirectional Encoder Representations from Transformers, Devlin et al. (2018)), and it is suitable to analyze financial text. Since it was trained on a rich dataset of financial corpus, FinBERT is shown to correctly account for the complex jargon of the financial domain, which generic models cannot capture accurately (Loughran and McDonald, 2011). For instance, the term "exposure" could refer to risk in one context or to the extent of investment in another. We provide more details on the class of BERT models in Appendix A.

3.3 Data

This section describes our data sources. We obtain the transcripts of the quarterly earnings calls from Bloomberg and Refinitiv, which we will use to compute sentiment. We extract a sentiment variable for the aggregate market, which we showed in the Introduction, using the online version of the *Wall Street Journal*. We obtain our financial and macro-economic variables from standard data providers.

3.3.1 Earnings Conference-Calls

We compute our measure of sentiment by analyzing the text of quarterly earnings call, which we obtain via Refinitiv and Bloomberg. We analyze them using a FinBERT model. Earnings calls are conducted by companies with their board members, investors, analysts and the press. These calls typically occur once every quarter and are used to discuss the company's financial results and performance. Earnings calls serve as a platform for the company to communicate its financial performance and outlook to stakeholders. They also provide an opportunity for investors and analysts to ask questions and gain insights into the company's operations and prospects.

Earnings calls usually follow a structured format, consisting of two parts: the main presentation and a question-and-answer (Q&A) session. During the main presentation portion, the company's management, typically the CEO and CFO, provide a detailed overview of the financial results for the quarter or year. Management may also discuss strategic initiatives, market trends, and other factors that have influenced the company's performance. The main presentation is scripted and prepared in advance to ensure that important information is communicated clearly and accurately. A Q&A session follows the main presentation. In this session, analysts, investors, and the press have the opportunity to interact with the company's management. These questions can cover a wide range of topics, including specific financial results, future guidance, industry trends, and more. The Q&A session allows for direct engagement between the company and its stakeholders, providing additional insights beyond the prepared remarks.

Additionally, earnings calls can offer insights into the company's sentiment and risk perceptions. This information is derived through textual analysis of transcripts from these calls. Earnings calls are considered valuable because they provide timely information on a company's financial performance and offer insights into its management's perspective on the business. Furthermore, they are used to monitor sentiment and risk trends, which can offer valuable insights into the overall health of a company and its industry.

3.3.2 News Sources

In our analysis, we occasionally rely on measures of sentiment extracted from business news. We consider two different sources. we estimate a sentiment measure from the online version of the *Wall Street Journal*, 1998:01 – 2022:04. We have around $3,000$ articles per month. We use the R package `sentimentr`, which accounts for valence shifters. We include an updated dictionary that combines Loughran and McDonald (2011) with a general purpose one. We also consider a measure of news sentiment from Shapiro et al. (2022), 1980:01 – 2023:01. The authors filter out the economic news from 16 major newspapers (including the NYT), via LexisNexis and use lexical methods with a combination of dictionaries.

3.3.3 Financial variables

As a representative sample of large and liquid stocks, we consider the companies in the Dow Jones Industrial Average index (DJIA). This gives us a cross-section of 30 assets, even though we will discard some of them due to lack of availability for the earnings calls. While our cross-section is not excessively large, we deem it to be representative. We estimate their common long-run trends starting from the 5 Fama-French factors (Fama and French, 2015).

When we consider the aggregate market, we obtain the S&P500 index together with its dividends from Robert Shiller's website. In order to estimate the cointegration system, we obtain data from the real potential Gross Domestic Product (GDP) from FRED, and data for the demographic variables used to compute the middle-aged to young ration from the U.S Census Bureau.

3.4 Results

This section reports our results. We describe how to obtain measures of deviations from trend for each asset. We later relate them to sentiment measures, and compare their performance in an error correction framework.

3.4.1 Long-Run Trends in Equity Markets

As a preliminary step to our empirical analysis, we need to compute a measure of asset-level mispricing. That is, we estimate a long-run stochastic trend, which is common in the cross-section of prices and we compute the deviation of prices from such estimated trend. We follow the approach in Favero et al. (2019), but we perform our analysis on stock prices, rather than portfolios.

Our cross-section includes the companies in the Dow Jones Industrial Average index. Following Section 3.2.1, we estimate the regression

$$p_{i,t} = \alpha_i t + \boldsymbol{\beta}_i' \log \boldsymbol{F}_t + u_{i,t}, \tag{3.6}$$

in which $p_{i,t}$ is the price of asset i at time t; t represents a linear deterministic trend; $\boldsymbol{F}_t$ collects the levels of buy-and-hold portfolios that replicate the Fama and French (2015) 5 factors; and $u_{i,t}$ is the residual from the regression.

$\boldsymbol{F}_t$ is the long-run trend, which is common across assets. We construct it starting from a set of factors. In particular, let $\boldsymbol{f}_t$ collect the return of the 5 Fama-French factors at time t. We build the long-run trend by cumulating such returns over quarters, i.e. $\log \boldsymbol{F}_t - \log \boldsymbol{F}_{t-1} = \boldsymbol{f}_t$. Notice that, even though the long-run drivers of such trend is common, the loadings $\boldsymbol{\beta}_i$ are asset specific. Also notice that $u_{i,t}$ is a measure of deviation from trend, which we interpret as mispricing, and it depends on both i and t. Such terms can only be recovered through price regressions.

Our sample goes from $1980 : Q1$ to $2023 : Q1$, at the quarterly frequency. We consider

28 out of 30 stocks in the DJIA due to data availability of their earnings call transcripts. In the following Sections, we work under the assumption that $u_{i,t}$ is stationary for each asset. We test the validity of such assumption by running cointegration tests. We want to see whether prices and long-run trends are cointegrated, and if $u_{i,t}$ is an Error Correction Term (ECT). For each asset in our cross-section, we perform an Augmented Dickey-Fuller test. For 18 out of 28 series, more than 60% of our assets, there is cointegration at the 10% level. This is in line with the findings in Favero et al. (2019), even though in this manuscript we consider single stocks, rather than portfolios. Interestingly, the cointegration relation still holds.

3.4.2 Sentiment vs Deviations From Trend

The first step in our analysis is to see whether measures of sentiment can be used as a proxy for the ECT in our error correction framework of Equation (3.3). We conduct an in-sample analysis. As a first evidence, we look at a simple panel regression, to see whether sentiment and ECT are correlated. That is, we estimate

$$u_{i,t} = \beta s_{i,t} + \alpha_i + \epsilon_{i,t}, \tag{3.7}$$

and we use a fixed effect model for α_i. $u_{i,t}$ is the ECT for company i in quarter t, while $s_{i,t}$ is the sentiment extracted with FinBERT from the earnings call transcript at time t. We have an unbalanced panel, due to the availability of the sentiment measure. Our data is at the quarterly frequency, and our estimation window goes from $2001 : Q1$ to $2023 : Q2$. We report our estimates in Table I.

Table I

ECT vs Sentiment

This table reports the coefficient estimates for the panel regression $u_{i,t} = \beta s_{i,t} + \alpha_i + \epsilon_{i,t}$. $u_{i,t}$ is the ECT for company i in quarter t, the residual from a regression of the log-price of asset i on risk-drivers. $s_{i,t}$ is the FinBERT sentiment extracted from the earnings call transcripts released by company i at time t. α_i includes firm fixed effects. Our unbalanced panel is made of the companies in the Dow Jones Industrial Average index (DJIA), and the sample period is $2001:Q1$ to $2023:Q2$, at the quarterly frequency. Standard errors are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively.

	$u_{i,t}$
$s_{i,t}$	0.215***
	(0.040)
N	2,169
R^2	0.013
Adjusted R^2	0.0004

The estimate of coefficient β from Equation (3.7) is 0.215, positive and statistically different from zero. This fact shows that the ECT and sentiment are positively correlated. That is, assets tend to be overpriced when their earnings statements are particularly positive. Of course, this is a simple correlation and we claim no causal relation between the two. The R^2 is quite low, as we do not aim at explaining the variations in $u_{i,t}$, which is affected by many other factors.

Having shown that the ECT and the sentiment are correlated, we now investigate whether the two variables share the same error correction properties. In particular, for each asset i in our cross-section we estimate the regressions

$$r_{i,t,t+1} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_{t+1} + \delta_i^u u_{i,t} + \epsilon_{i,t+1}, \tag{3.8}$$

$$r_{i,t,t+1} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_{t+1} + \delta_i^s s_{i,t} + \epsilon_{i,t+1}, \tag{3.9}$$

and compare the coefficients δ_i^u and δ_i^s. We keep the same estimation window as for Table I, i.e. from $2001:Q1$ to $2023:Q2$. Even though we could estimate Equation (3.8) using a longer time series, we restrict the sample for the sake of comparability. We present our results in Figures 3.2 and 3.3.

Figure 3.2 reports the estimates of δ_i^u for each company i, obtained by estimating Equation (3.8). The horizontal bars represent the 90% confidence interval. All the coefficients are negative, in line with our error correction framework. Since $\delta_i^u < 0$, a price higher than what implied by its long-run trend, such that $u_{i,t} > 0$, will have a lower return in the following quarter. The deviations from equilibrium are thus corrected over time. Most of the coefficients are statistically different from 0. Interestingly, while Favero et al. (2019) run the same estimation procedure in a cross-section of portfolios, we obtain comparable results with single stocks, even though with a larger degree of noise. Figure

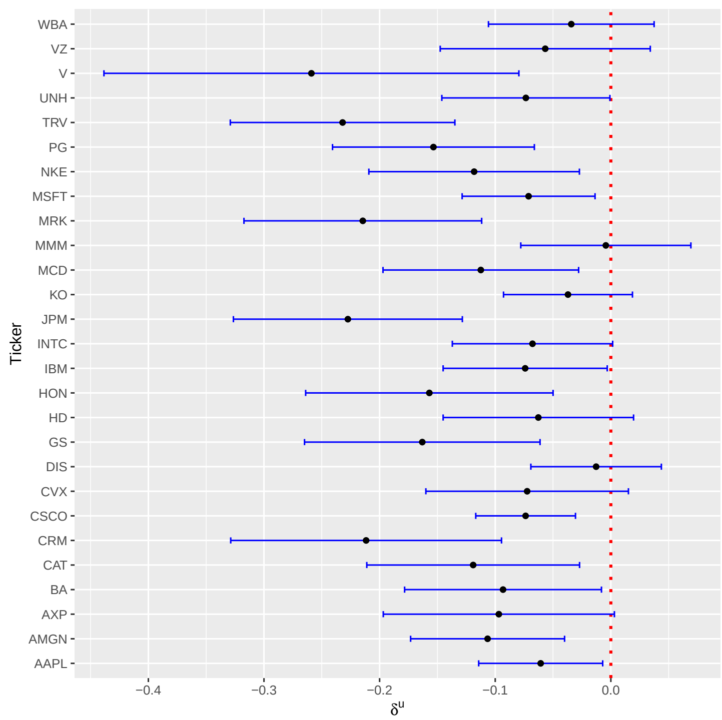


Figure 3.2: The picture depicts the estimated coefficients δ_i^u in the regression $r_{i,t,t+1} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_t + \delta_i^u u_{i,t} + \epsilon_{i,t+1}$, for each company i in the Dow Jones Industrial Average. The horizontal bars represent the 90% confidence interval.

3.3 reports the estimates of δ_i^s for each company i, obtained by estimating Equation (3.9), together with their 90% confidence interval. Interestingly, in this case no coefficient is statistically different from 0. The sentiment measure by itself does not predict future returns. Moreover, even if we neglected the error bars, the δ_i^s estimates are positive for some assets. This fact is not consistent with an error correction mechanism. To sum up,

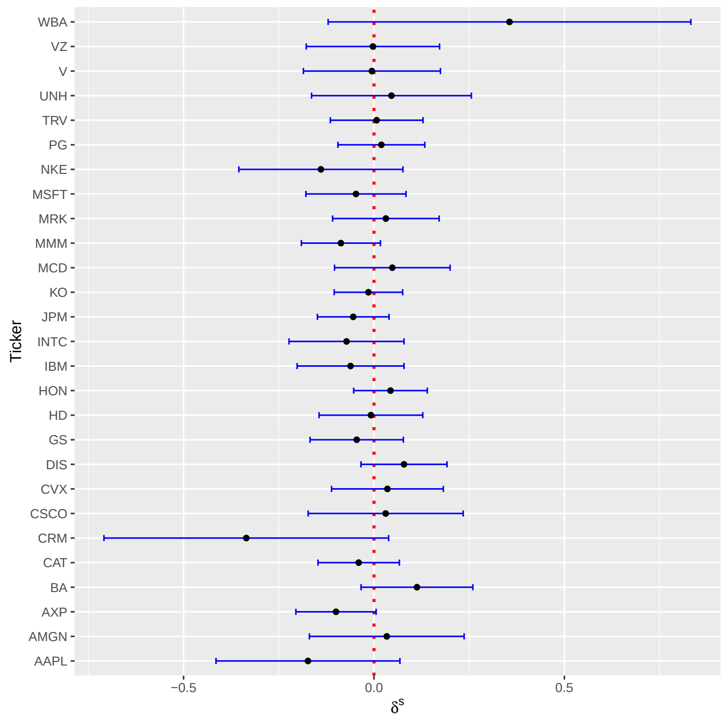


Figure 3.3: The picture depicts the coefficients δ_i^s in the regression $r_{i,t,t+1} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_t + \delta_i^s s_{i,t} + \epsilon_{i,t+1}$, for each company i in the Dow Jones Industrial Average. The horizontal bars represent the 90% confidence interval.

we showed that measures of sentiment obtained from earnings calls and mispricings are positively related. While the ECTs can be used as predictors for the assets' future returns, sentiment appears to have no predictive power. Sentiment by itself does not fit in a simple error correction framework. Therefore, we cannot reliably use sentiment as a proxy for asset mispricing. This fact may be due to the larger variance of our sentiment measure compared to the ECTs. What the ECT captures is not included in the sentiment. In the next sections we explore whether sentiment can still enter an error correction framework, by affecting the speed at which the equilibrium is restored in a system.

3.4.3 Speed of Adjustment

The evidence in the previous section leads us to conclude that sentiment measures cannot be directly employed as a proxy for mispricings. In this section, we show that such a measure can enter the error correction framework through the speed of adjustment. The speed of adjustment is defined as the coefficient in front of the error correction term, e.g. δ_i in Equation (3.3). The larger the coefficient, the faster the price gets back to its long-run trend. Favero et al. (2019) assume that this coefficient is constant over time, and we followed this approach in Section 3.4.2.

We now investigate whether the speed of adjustment is time-varying, and if sentiment affects it. To understand why sentiment may affect the time to get back to equilibrium, consider a behavioral model of diagnostic expectations as in Bordalo et al. (2019) and Bordalo et al. (2021). When agents receive positive news about a company, they overreact and increase their demand for it, which drives the price upwards. Over time, agents learn about the true new fundamental value of the asset, and trade so that its price gets back to equilibrium. However, the speed to which they update their beliefs may be affected by sentiment. Receiving positive signals from an earnings call confirms their beliefs about the higher price, which implies that prices will take more time to get back to equilibrium.

We develop an econometric model in which the speed of adjustment is a linear function of sentiment. Since we are also interested in the forecasting and asset allocation properties of such specification, we include the earnings call released in the same quarter as the ECT. For each asset i, we estimate the regression

$$r_{i,t,t+1} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_{t+1} + (\delta_{0,i} + \delta_{1,i} s_{i,t})\, u_{i,t} + \epsilon_{i,t+1}. \tag{3.10}$$

We have already discussed that we expect the overall coefficient in front of $u_{i,t}$ to be negative, in line with an error correction framework. Therefore, we should have $\delta_{0,i} < 0$. If sentiment really affects the speed of adjustment as in our heuristic argument, we expect

$\delta_{1,i} > 0$. Indeed, suppose that the i-th asset is overpriced, $u_{i,t} > 0$. If the sentiment is positive, it will lead to a lower speed of adjustment, i.e. the overpricing will last for longer. If a positive sentiment needs to lower the absolute value of the coefficient, we will have $\delta_{1,i} > 0$. We present the estimates for $\delta_{1,i}$ from Equation (3.10) in Figure 3.4. For most

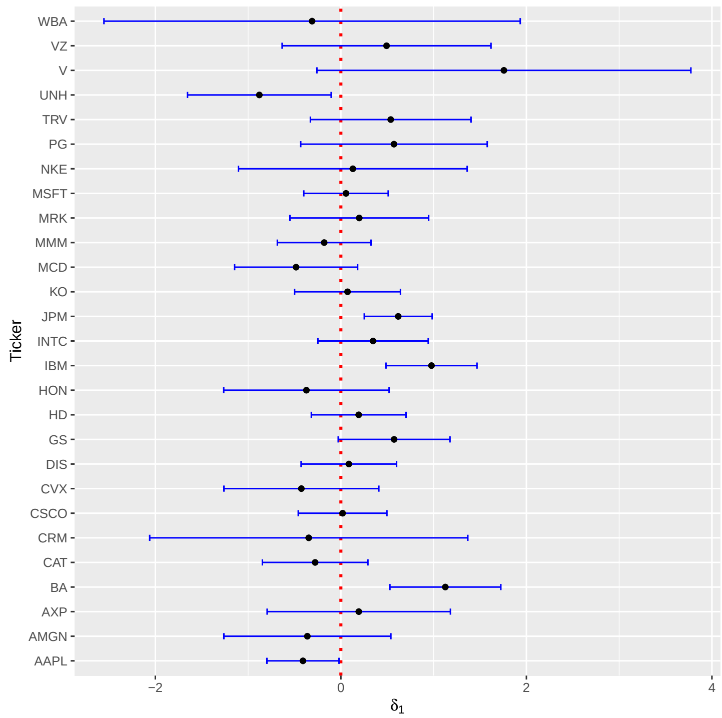


Figure 3.4: The picture depicts the coefficients $\delta_{1,i}$ in the regression $r_{i,t,t+1} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_t + (\delta_{0,i} + \delta_{1,i} s_{i,t}) u_{i,t} + \epsilon_{i,t+1}$, for each company i in the Dow Jones Industrial Average. The horizontal bars represent the 90% confidence interval.

companies, the coefficient $\delta_{1,i}$ appears to be positive but not statistically significant, due to the large error bars. However, in the literature there is evidence of asymmetric reaction to shocks, so we separate the speed of adjustment in the cases where $u_{i,t}$ is positive and negative. In particular, our preferred approach is to estimate the regression

$$r_{i,t,t+1} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_{t+1} + \left(\delta_{0,i}^{+} + \delta_{1,i}^{+} s_{i,t}\right) u_{i,t}^{+} + \left(\delta_{0,i}^{-} + \delta_{1,i}^{-} s_{i,t}\right) u_{i,t}^{-} + \epsilon_{i,t+1}, \qquad (3.11)$$

in which $u_{i,t}^{+} = u_{i,t} 1 (u_{i,t} > 0)$ and $u_{i,t}^{-} = u_{i,t} 1 (u_{i,t} < 0)$. The specification in Equation (3.11) lets us separate the speed of adjustment dynamics. We show the estimates of δ_1^{+}

in Figure 3.5, and the ones for δ_1^- in Figure 3.6 . The two figures display an asymmetric

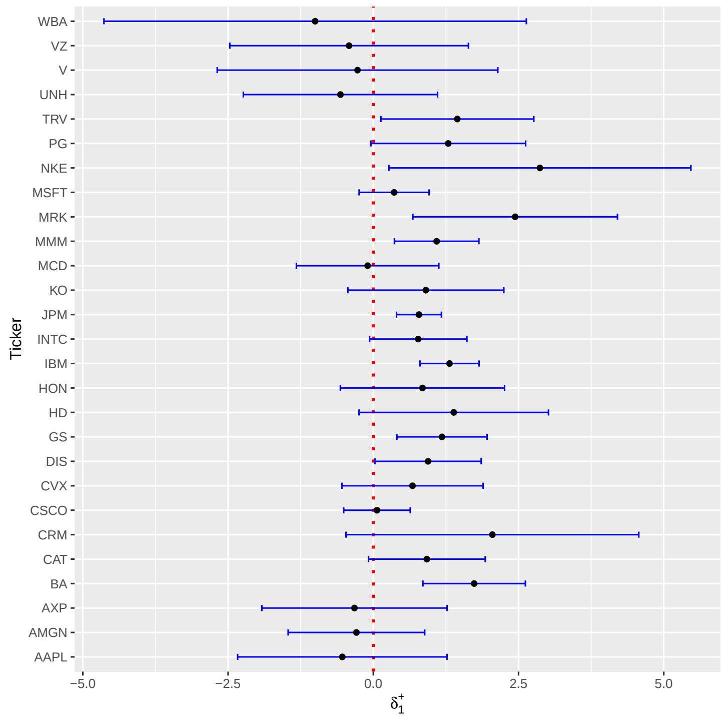


Figure 3.5: The picture depicts the coefficients $\delta_{1,i}^+$ in the regression $r_{i,t,t+1} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_t + \left(\delta_{0,i}^+ + \delta_{1,i}^+ s_{i,t}\right) u_{i,t}^+ + \left(\delta_{0,i}^- + \delta_{1,i}^- s_{i,t}\right) u_{i,t}^- + \epsilon_{i,t+1}$, for each company i in the Dow Jones Industrial Average. The horizontal bars represent the 90% confidence interval.

behavior. In Figure 3.5, most of the δ_1^+ coefficients are positive. Even though the error bars are large, a relevant number of estimates is statistically different from zero, as we would expect from our heuristic reasoning. When prices are above their long-run trend, positive sentiment reduces the speed of adjustment, while negative sentiment makes the disequilibrium disappear faster. The width of the confidence intervals may be related to the relatively low number of observations included in each separate regression. We will soon tackle this issue by imposing restrictions on the coefficients.

A different picture emerges from Figure 3.6. Most of the δ_1^- estimates are not statistically different from zero and, if anything, most of them are negative. This evidence is consistent with a model with asymmetric reaction to news. In particular, even though an asset is under-priced, the arrival of positive news in the form of sentiment does not

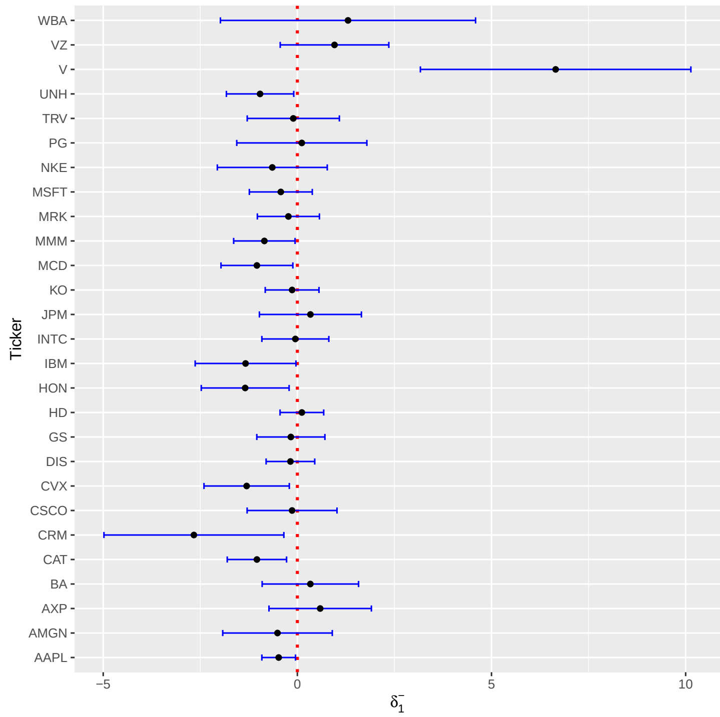


Figure 3.6: The picture depicts the coefficients $\delta_{1,i}^-$ in the regression $r_{i,t,t+1} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_t + \left(\delta_{0,i}^+ + \delta_{1,i}^+ s_{i,t}\right) u_{i,t}^+ + \left(\delta_{0,i}^- + \delta_{1,i}^- s_{i,t}\right) u_{i,t}^- + \epsilon_{i,t+1}$, for each company i in the Dow Jones Industrial Average. The horizontal bars represent the 90% confidence interval.

dramatically affect the speed to go back to equilibrium.

We saw that most of the estimates for δ_1^+ and δ_1^- have a similar magnitude. However, their error bars are quite large, given that we estimate them for each asset, and we necessarily have a small number of observations. We fix this issue by estimating a system of equations and imposing restrictions on the coefficients. In particular, we will assume that the speed of adjustment parameters are common across stocks. We are thus assuming that, irrespective of the type of stock considered, given a level of mispricing, the path towards equilibrium is comparable. This assumption can be defended in our case, as we are considering some of the most liquid in the New York Stock Exchange (NYSE), as they are all included in the DJIA. We estimate the system of equations

$$r_{i,t,t+1} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_{t+1} + \left(\delta_0^+ + \delta_1^+ s_{i,t}\right) u_{i,t}^+ + \left(\delta_0^- + \delta_1^- s_{i,t}\right) u_{i,t}^- + \epsilon_{i,t+1}, \tag{3.12}$$

in which we have imposed $\delta^{\pm}_{0,i} = \delta^{\pm}_0$ and $\delta^{\pm}_{1,i} = \delta^{\pm}_1$. We present the estimates the estimates of the restricted coefficients of Equation (3.12) in Table II.

Table II
Speed of Adjustment

This table reports the coefficient estimates for the regression $r_{i,t,t+1} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_{t+1} + \left(\delta_0^+ + \delta_1^+ s_{i,t}\right) u^+_{i,t} + \left(\delta_0^- + \delta_1^- s_{i,t}\right) u^-_{i,t} + \epsilon_{i,t+1}$. $r_{i,t,t+1}$ is the log return of asset i between quarter t and $t+1$. $\boldsymbol{f}_{t+1}$ collects the log-returns of the 5 Fama-French factors. $u_{i,t}$ is the ECT for company i in quarter t, the residual from a regression of the log-price of asset i on risk-drivers; $u^+_{i,t}$ ($u^-_{i,t}$) denotes its positive (negative) part. $s_{i,t}$ is the FinBERT sentiment extracted from the earnings call transcripts released by company i at time t. We include the companies in the Dow Jones Industrial Average index (DJIA), and the sample period is $2001:Q1$ to $2023:Q2$, at the quarterly frequency. Standard errors are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively.

	δ_0^+	δ_1^+	δ_0^-	δ_1^-
$r_{i,t,t+1}$	-0.17^{***}	0.19^{*}	-0.09^{***}	0.03
	(0.04)	(0.10)	(0.03)	(0.09)

Both the estimates of δ_0^+ and δ_0^- are negative and statistically different from 0 and from each other. Notice that we have not demeaned our sentiment measure, but our results are not affected by this choice (except for the magnitude of the coefficients). Indeed, the average sentiment in our sample is positive, as it is made by statements by the top managers, who tend to speak positively about their performance. We notice that δ_1^+ is positive statistically different from zero, in line with our findings in Figure 3.5. On the contrary, δ_1^- is smaller by an order of magnitude, and it is not statistically different from zero.

3.4.4 Symmetry Tests

In the previous section, we showed that sentiment affects the speed of adjustment, provided that we separate between positive and negative mispricings. By looking at the results in Table II, we can see that there is a clear asymmetry between the two cases.

However, in this section we show that the speed of adjustment is symmetric unconditionally, but asymmetric conditionally.

We start by testing the null hypothesis:

$$H_0 : \delta_0^+ + \delta_1^+ \mathbb{E}\left[S_{i,t}\right] = \delta_0^-, \quad \forall i. \tag{3.13}$$

Equation (3.13) tests whether the speed of adjustment is symmetric unconditionally. We are not explicitly including δ_0^- in the H_0, as we see from Table II that the coefficient is not statistically different from zero, and it is smaller by one order of magnitude. We verified that including it in the hypothesis does not affect the results. We test the null hypothesis by using an F-test, which amounts to testing a set of joint restrictions on the coefficients of a system of equations. We get a p value which is close to 1, therefore we clearly cannot reject the null. We conclude that the speed of adjustment is symmetric conditionally with respect to the sign of the deviation from the equilibrium.

Equivalently, another test of unconditional symmetry can be obtained by estimating the regression

$$r_{i,t,t+1} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_{t+1} + \delta^+ u_{i,t}^+ + \delta^- u_{i,t}^- + \epsilon_{i,t+1}. \tag{3.14}$$

Equation (3.14) imposes the restrictions that the coefficients in front of the error correction terms, δ^+ and δ^-, are constant in the cross-section. Of course, these coefficients represent a constant speed of adjustment. We report the results in Table III. In line with the F-test of Equation (3.13), we can see that the speed of adjustment is uncoditionally symmetric, as the two coefficients δ^+ and δ^- are clearly within each other's 90% confidence intervals.

On the contrary, the speed of adjustment is conditionally asymmetric. We can look at the time evolution of the speed, i.e. at the behavior of $\left(\delta_{0,i}^+ + \delta_{1,i}^+ s_{i,t}\right)$ vs $\left(\delta_{0,i}^- + \delta_{1,i}^- s_{i,t}\right)$ for each company in our sample. We reject that the two objects are conditionally the same. As an example, Figure 3.7 shows the time varying coefficient for AAPL (Apple Inc) and AXP (American Express Co.). Figure 3.7 clearly shows that the "overpricing" speed of

Table III
Speed of Adjustment - Unconditional

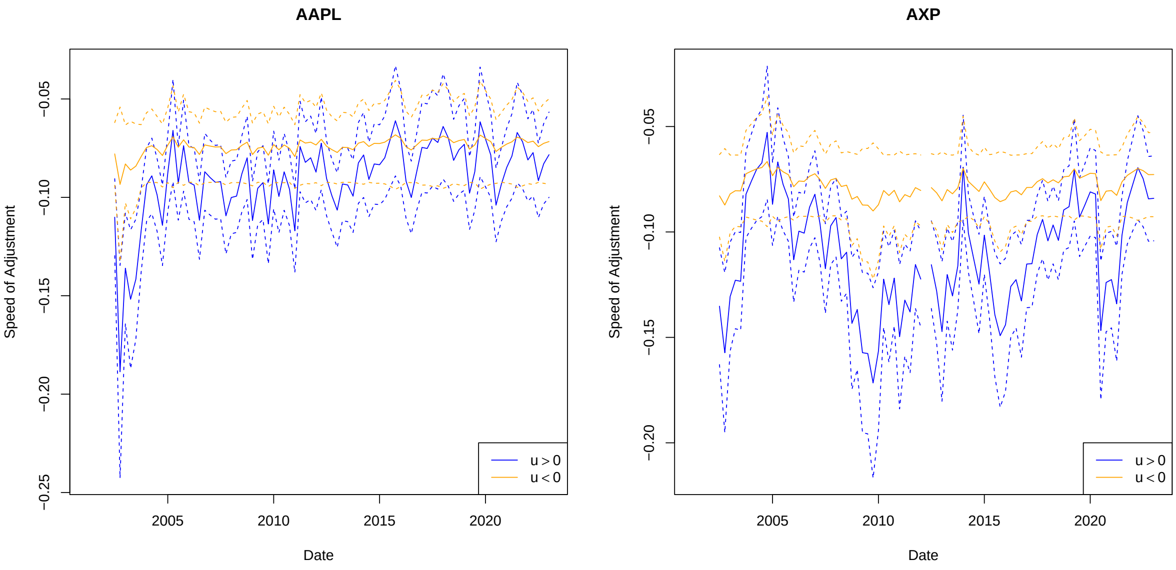
This table reports the coefficient estimates for the regression $r_{i,t,t+1} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_{t+1} + \delta^+ u_{i,t}^+ + \delta^- u_{i,t}^- + \epsilon_{i,t+1}$. $r_{i,t,t+1}$ is the log return of asset i between quarter t and $t+1$. $\boldsymbol{f}_{t+1}$ collects the log-returns of the 5 Fama-French factors. $u_{i,t}$ is the ECT for company i in quarter t, the residual from a regression of the log-price of asset i on risk-drivers; $u_{i,t}^+$ ($u_{i,t}^-$) denotes its positive (negative) part. We include the companies in the Dow Jones Industrial Average index (DJIA), and the sample period is 2001 : $Q1$ to 2023 : $Q2$, at the quarterly frequency. Standard errors are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively.

	δ^+	δ^-
$r_{i,t,t+1}$	-0.094^{***}	-0.080^{***}
	(0.018)	(0.016)

adjustment (i.e. when $u_{i,t} > 0$) is statistically different from the "underpricing" one (i.e. when $u_{i,t} < 0$). The difference between the two series varies over time. The result seems to be driven by the periods of particularly negative sentiment, as both δ_1^+ and δ_1^- are positive (therefore, a lower sentiment is associated with a higher speed of adjustment, in absolute value).

3.5 Conclusion

In this chapter, we investigate whether measures of sentiment extracted from quarterly earnings conference-calls affect the dynamics of stock prices. Using a cross-section of publicly traded companies, we show that sentiment positively correlates with price deviations from their long-run trend, estimated via an error correction model. We document that even though sentiment does not predict future stock returns, it impacts the speed at which prices revert to equilibrium. We find asymmetric effects on overpriced and underpriced stocks. Potential future steps could include increasing our cross-section, and formally deriving a model that explains the asymmetry behind our results. Moreover, we should investigate the determinants of predictability in our framework, potentially in line with



(a) Apple Inc **(b)** American Express Co.

Figure 3.7: The picture depicts the speed of adjustment for Apple Inc (panel (a)) and American Express Co. (panel (b). The blue (orange) solid line represent the speed of adjustment for a positive (negative) mispricing, $\left(\delta^+_{0,i} + \delta^+_{1,i} s_{i,t}\right)$ $\left(\left(\delta^-_{0,i} + \delta^-_{1,i} s_{i,t}\right)\right)$. The dotted lines represent the 90% confidence intervals.

the market and volatility timing decomposition approach of Goulding et al. (2023).

A BERT (Bidirectional Encoder Representations from Transformers)

An important advance in the class of Natural Language Processing (NLP) models has been the introduction of BERT (Bidirectional Encoder Representations from Transformers), a model developed by Devlin et al. (2018). BERT's architecture and training methodology represent a substantial advancement in the field, enabling improved and unmatched performances on a wide range of NLP tasks. BERT is built upon the Transformer model introduced by Vaswani et al. (2017). The Transformer model eschews conventional recurrent or convolutional layers, focusing instead on self-attention mechanisms to process text. Figure 8, which is take from Devlin et al. (2018) illustrates the architecture of a simple BERT model. BERT's training involves two stages: pre-training and fine-tuning. The

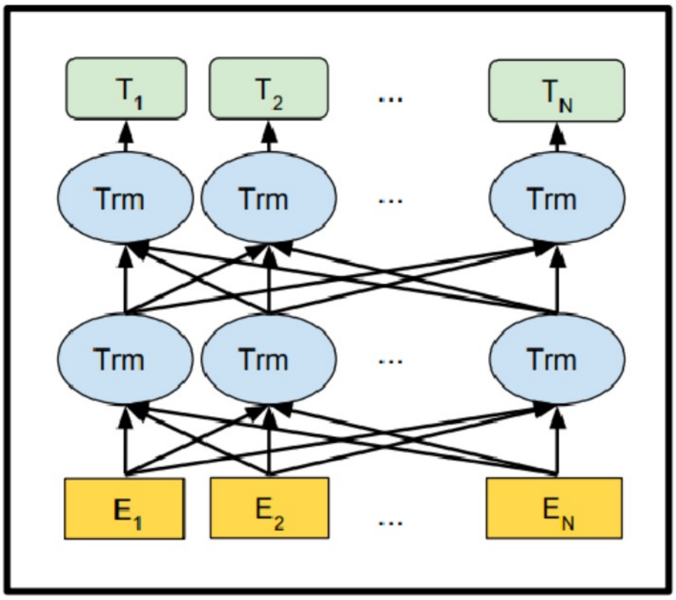


Figure 8: The picture depicts the BERT model architecture.

pre-training stage is unsupervised and utilizes two novel tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). The fine-tuning stage is conducted after the pre-training, and fine-tunes BERT for specific tasks, wherein the entire model is slightly adjusted. This stage requires significantly less data compared to training a model from scratch. Figure 9, which is taken from Devlin et al. (2018) depicts the pre-training and fine-tuning stages. BERT's input representation is a blend of WordPiece

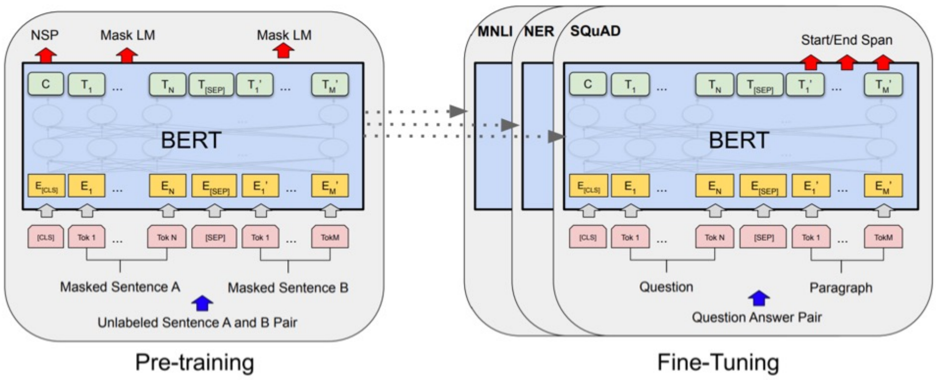


Figure 9: The picture depicts the pre-training and fine-tuning stages for BERT.

token embeddings, positional embeddings, and segment embeddings. This approach allows BERT to handle out-of-vocabulary words effectively and provides the model with necessary positional and contextual information.

FinBERT's fine-tuning on specialized NLP tasks has resulted in performances that surpass those of traditional machine learning models, deep learning alternatives, and even fine-tuned versions of the original BERT model. Each fine-tuned variant of FinBERT is designed for a specific purpose and is readily accessible to the public via the Huggingface platform. FinBERT's pre-training encompasses a vast corpus of financial communication texts, amounting to 4.9 billion tokens. This corpus includes 2.5 billion tokens from Corporate Reports (10-K & 10-Q), 1.3 billion tokens from Earnings Call Transcripts, and 1.1 billion tokens from Analyst Reports.

B MIDAS

Our main dataset and empirical strategy is at the quarterly frequency, due the release schedule of the earnings calls. However, we potentially have more text information at higher frequency from business news. In this Appendix, we see whether this information can be useful in the context of forecasting and asset allocation, within a long-run error correction model. Even though we cannot conduct this analysis for each company in our sample, as some of them are overly represented in the news, we can do so for the overall market.

As we discussed above, sentiment measures are available at a higher frequency than the ECT. We want to test whether they help in forecasting the market, on top of what the ECT already does. Since we have variables at different frequencies, we use a mixed data sampling (MIDAS) approach (Ghysels et al. (2004), Armesto et al. (2010)). In particular, we will compare regressions in the form

$$r_{t,t+1} = \alpha + \delta u_t + \epsilon_{t+1}, \tag{15}$$

$$r_{t,t+1} = \alpha + \delta u_t + \gamma_1 s_{t+1/3} + \epsilon_{t+1}, \tag{16}$$

$$r_{t,t+1} = \alpha + \delta u_t + \gamma_1 s_{t+1/3} + \gamma_2 s_{t+2/3} + \epsilon_{t+1}, \tag{17}$$

with $s_{t+k/3}$ denoting the sentiment at the end of the k-th month of quarter t. We add the new sentiment information as it becomes available, and we forecast the same one-quarter-ahead market return over the three specifications.

A potential problem in the MIDAS framework comes from the large number of high-frequency coefficients that need to be estimated. The literature usually imposes a parametric structure on the coefficients (Armesto et al., 2010). Since we have a low number of variables, we estimate all the coefficients without imposing any restrictions, which helps with interpretation. In the case of long-horizon forecasts we will impose the functional form

$$\gamma_k = \gamma_k(\theta) = \gamma \frac{\exp\{(\theta k^2)\}}{\sum_{j=1}^{J} \exp\{(\theta j^2)\}}, \tag{18}$$

with θ being a hyperparameter and J being the total number of high-frequency lags.

B.1 Out-of-Sample

We compare the models discussed above in an out-of-sample exercise. We evaluate the forecasting performance out-of-sample using the the Campbell and Thompson (2008) out-of-sample R^2, henceforth R^2_{OS}. The R^2_{OS} statistics compares the forecast errors of a candidate model to a benchmark. In particular,

$$R^2_{OS} = 1 - \frac{MSFE_1}{MSFE_0}, \tag{19}$$

in which $MSFE_1$ ($MSFE_0$) is the Mean Squared Forecast Errors of the candidate (benchmark) model. $R^2_{OS} > 0$ implies that the candidate model has a superior forecasting performance than the benchmark. We follow the asset pricing literature and we use the the trailing mean of the market excess returns as benchmark model (Welch and Goyal, 2008).

We test the null hypothesis $H_0 : R^2_{OS} \leq 0$ against the alternative $H_A : R^2_{OS} > 0$ using the Clark and West (2007) procedure. Let $\hat{r}^0_{t-1,t}$ and $\hat{r}^1_{t-1,t}$ be the forecasts of the market return under the benchmark and candidate model respectively. The forecast errors $\hat{e}_{0,t}$ and $\hat{e}_{1,t}$ can be obtained as

$$\hat{e}_{j,t} = r_{M,t-1,t} - \hat{r}^j_{M,t-1,t}, \quad j = 0, 1. \tag{20}$$

Clark and West (2007) propose to fit the regression

$$\hat{e}^2_{0,t} - \hat{e}^2_{1,t} + \left(\hat{r}^0_{M,t-1,t} - \hat{r}^1_{M,t-1,t}\right)^2 = \mu + \epsilon_t \tag{21}$$

via OLS and to do a t-test of $H_0 : \mu \leq 0$ against $H_A : \mu > 0$. Since $\mu \leq 0$ is equivalent to $MSFE_0 \leq MSFE_1$, this procedure let us test $H_0 : R^2_{OS} \leq 0$. We compute t-statistics using heteroskedasticity and autocorrelation consistent (HAC) standard errors (Newey and West, 1986).

As a final remark, we will inspect not only the statistical significance of the R^2_{OS}, but also its magnitude. Campbell and Thompson (2008) show that a monthly R^2_{OS} of 0.5% is already economically significant. Therefore, we will always compare our estimates with this threshold.

B.2 In-Sample

We run an in-sample analysis, simply comparing the time series of sentiment and and the ECT. The first step is to see whether our claim that we can use the former as a proxy for the latter is reflected in the data. We start by looking at the aggregate market index. We do so, because we have to consider only one series, and only later will we go on to consider the whole cross-section. We can thus see graphical evidence.

We report our results in Figure 3.1. The left panel (a) of the figure depicts a sentiment measure computed from the headlines of the *Wall Street Journal*. We can see that, when we have both the ECT and the sentiment, the two series strikingly move together, with a correlation of 0.52. This result is even more intriguing, as they are computed from two totally separated datasets and procedures.

Even though the WSJ sentiment highly correlates with the ECT, it is hard to use it in an out-of-sample analysis due to limited data availability. Indeed, it is available for roughly 25 years. Therefore, for our main analysis we will resort to the sentiment indicator computed by Shapiro et al. (2022), who use several newspaper sources for economic news over the whole span of our sample. The correlation between such indicator and the ECT drops to 0.34, which is still high. The results are depicted in the right panel (b) of the figure. It is clear that the two sentiment series track each other, which is reassuring.

Table IV

In-Sample MIDAS

This table reports the coefficient estimates from the MIDAS regression $r_{t,t+1} = \alpha + \boldsymbol{\beta}'\boldsymbol{x}_t + \epsilon_{t+1}$. $r_{t,t+1}$ is the market log return between quarter t and $t+1$. $\boldsymbol{x}_t$ contains different regressors at different frequencies. u is the ECT, the residual from a regression of market log prices on risk-drivers. $s_{t+k/3}$ is the sentiment at the end of the k-th month of quarter t, from Shapiro et al. (2022). $\Delta s_{t+k/3,t+j/3}$ is the sentiment difference between months j and k of quarter t. We use quarterly observations for r and u, and monthly data for s. p values are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively. The sample period is $1980Q1 : 2022Q3$.

	$r_{t,t+1}$	$r_{t,t+1}$	$r_{t,t+1}$	$r_{t,t+1}$
u_t	-0.23^{*}	-0.27^{***}	-0.23^{**}	-0.23^{**}
	(0.05)	(0.00)	(0.04)	(0.04)
$s_{t+2/3}$		0.80^{***}		
		(0.00)		
$s_{t+1/3}$		-0.68^{***}		
		(0.00)		
$\Delta s_{t+1/3,t+2/3}$			0.30^{***}	
			(0.00)	
$\Delta s_{t,t+1/3}$			0.07	-0.22^{*}
			(0.32)	(0.06)
$\Delta s_{t,t+2/3}$				0.46^{***}
				(0.00)
α	0.25^{***}	0.19^{**}	0.20^{***}	0.20^{***}
	(0.00)	(0.04)	(0.01)	(0.01)
N	170	170	170	170
R^2	0.06	0.16	0.16	0.16
Adjusted R^2	0.05	0.15	0.14	0.14

After establishing such correlation between sentiment and ECT, we want to use the former as a proxy for the latter in a forecasting exercise. The key point is that the two series are available at a different frequency, and as we already mentioned we will use a simple MIDAS framework. We run the regressions in Equations (15) and (17), and we report our results in Table IV. Our sample goes from $1980Q1$ to $2022Q3$.

We can notice that the coefficient on u_t is always negative and significant. This is in line with Favero et al. (2019), who claim that this variable is indeed an error correction term. Moreover, the coefficient is fairly stable across specifications, and it stays significant as we add the sentiment measures. We notice that the adjusted R^2 dramatically increases as we add the sentiment measures to the regression, going from 5% to 15%.

We can also see that, when estimating Equation (17), the coefficients on $s_{t+2/3}$ and $s_{t+1/3}$ have comparable magnitudes, but opposite signs. This motivates us to run a regression using variations in sentiment, $\Delta s_{t+1/3,t+2/3} = s_{t+2/3} - s_{t+1/3}$ as the independent variable. By doing so, we impose that the two coefficients have the same magnitude. We report the results in column 3. The coefficient on this new variable is still positive and highly significant. As expected, the adjusted R^2 is still above the specification with the simple ECT.

Finally, we run a regression in which we use $\Delta s_{t,t+1/3}$ and $\Delta s_{t,t+2/3}$ as independent variables, we report the results in column 4. These indicators represent the cumulative changes in sentiment from the start of the quarter to the end of the k-th month. Once again, the coefficient on the ECT is not dramatically affected, the R^2 is still higher than in the first specification, and the sentiment coefficients are significant.

We increase the forecasting horizon by predicting the market return two quarters ahead. We present our estimates in Table V. We report our results from Table IV in the first column for convenience.

We can notice that the ECT is even more negative and significant at a longer horizon. This is in line with the results of Favero et al. (2019), and with the cointegration literature. As we add the sentiment measures in column 3, we see that the adjusted R^2 increases from 12% to 23%. Most of the sentiment estimates are insignificant. Most likely, this is due to not restricting the high-frequency coefficients, which we do for the ease of interpretability (Armesto et al., 2010).

B.3 Out-of-Sample

This section presents our out-of-sample results. In our analysis, we estimate recursively Equations (15) to (17), and we forecast the market return over the whole quarter. The

Table V

In-Sample MIDAS - Longer Horizon

This table reports the coefficient estimates from the MIDAS regression $r_{t,t+2} = \alpha + \boldsymbol{\beta}'\boldsymbol{x}_t + \epsilon_{t+1}$. $r_{t,t+2}$ is the market log return between quarter t and $t+2$. $\boldsymbol{x}_t$ contains different regressors at different frequencies. u is the ECT, the residual from a regression of market log prices on risk-drivers. $s_{t+k/3}$ is the sentiment at the end of the k-th month of quarter t, from Shapiro et al. (2022). We use quarterly observations for r and u, and monthly data for s. p values are shown in parentheses. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively. The sample period is $1980Q1 : 2022Q3$.

	$r_{t,t+1}$	$r_{t,t+2}$	$r_{t,t+2}$
u_t	-0.27^{**}	-0.33^{**}	-0.36^{***}
	(0.00)	(0.01)	(0.00)
$s_{(t+1)+2/3}$			0.56^{***}
			(0.00)
$s_{(t+1)+1/3}$			-0.02
			(0.95)
$s_{(t+1)}$			-0.30
			(0.25)
$s_{t+2/3}$	0.80^{***}		0.29
	(0.00)		(0.18)
$s_{t+1/3}$	-0.68^{***}		-0.13
	(0.00)		(0.65)
s_t			-0.16
			(0.33)
α	0.19^{**}	0.32^{**}	0.25^{**}
	(0.04)	(0.01)	(0.03)
N	170	169	169
R^2	0.16	0.12	0.26
Adjusted R^2	0.15	0.12	0.23

spirit of this exercise is to progressively add new information, and to see whether it improves our forecasts.

We present estimates for the Campbell and Thompson (2008) R^2_{OS} int Table VI. We always use the historical mean as a benchmark, in line with Welch and Goyal (2008). Our out-of-sample period is 2001Q2 to 2022Q3, and we use recursive estimates. $\mathbb{E}\left[r_{t,t+1} \mid I_{t+k/3}\right]$ means that the forecast for the market return over quarter t and $t+1$ is constructed in the k-month of quarter t.

Table VI
R^2_{OS} Statistics

This table reports the Campbell and Thompson (2008) out-of-sample R^2 (R^2_{OS}) in percent for market log returns forecasts between quarter t and $t+1$, at different months within the quarter. $\mathbb{E}\left[r_{t,t+1} \mid I_{t+k/3}\right]$ means that the forecast is constructed in the k-month of quarter t. The benchmark model is the historical mean. Recursive estimates. The out-of-sample period is 2001Q2 : 2022Q3. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively for the positive R^2_{OS} based on the Clark and West (2007) test.

	$\mathbb{E}\left[r_{t,t+1} \mid I_t\right]$	$\mathbb{E}\left[r_{t,t+1} \mid I_{t+1/3}\right]$	$\mathbb{E}\left[r_{t,t+1} \mid I_{t+2/3}\right]$
R^2_{OS}	1.98	−0.15	8.77***
	(0.43)	(0.60)	(0.01)

We notice that already the first column has an R^2_{OS} greater than 0.5%, which is the threshold for economic significance as of Campbell and Thompson (2008). However, this estimate is not statistically significant. This is to be expected, as we are not using the fully-fledged model of Favero et al. (2019), but a reduced version in which we only consider the previous-quarter ECT. That is, the forecasting horizon is rather short, and we replace the $\Delta \boldsymbol{x}_{t,t+1}$ factors of Equation (3.3) with 0.

However, we can notice that when we use information at the end of the second month in the quarter, the R^2_{OS} dramatically increases, reaching 8.77%. Not only we get a better forecasting power than for the ECT-only specification, but we improve with respect to the historical average. This result is both highly economically and statistically significant. Instead, when we only consider the first quarter sentiment, we get a negative R^2_{OS}, i.e.

this model performs worse than the historical benchmark. This could be related to the fact that we are using the ECT at time t, and not another sentiment measure s_t.

So far, we presented evidence in favor of using sentiment as a high-frequency predictor. We showed that adding it improves the forecasting power out-of-sample. We complement our analysis by implementing a Pesaran and Timmermann (1992) sign test. That is, we test whether each model helps predicting, if not the magnitude, at least the direction of future market returns. We present the p values from such test in Table VII.

Table VII

Pesaran-Timmermann

This table reports the p values for the Pesaran and Timmermann (1992) sign test. $\mathbb{E}\left[r_{t,t+1} \mid I_{t+k/3}\right]$ means that the forecast is constructed in the k-month of quarter t. Recursive estimates. The out-of-sample period is $2001Q2 : 2022Q3$. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively

	$\mathbb{E}\left[r_{t,t+1} \mid I_t\right]$	$\mathbb{E}\left[r_{t,t+1} \mid I_{t+1/3}\right]$	$\mathbb{E}\left[r_{t,t+1} \mid I_{t+2/3}\right]$
p	0.22	0.05*	0.00***

We can see that, even though when using the ECT alone we cannot reject the null hypothesis, we can do so whenever we add the sentiment measures. That is, sentiment helps predicting the direction of the market. A natural next step is to perform an asset allocation exercise in the Pesaran and Timmermann (1992) spirit: we invest the whole portfolio in the market whenever we have a positive forecasts, and in the risk-free rate otherwise.

B.4 Asset Allocation

We perform an asset allocation exercise. In each quarter t, you adjust your portfolio with an investment horizon until quarter $t+1$. In each month, (so, at the beginning, at $t+1/3$ and $t+2/3$), you add a $1/3$ weight to the market portfolio, if you forecast a positive excess return, and you invest $1/3$ less in the risk-free asset. Otherwise, you subtract $1/3$, and invest the additional weight in the risk free rate. Weights must always be between 0 and 1, and we impose that the initial portfolio is all invested in the market. The benchmark is the market portfolio. We include our results in Figure 10.

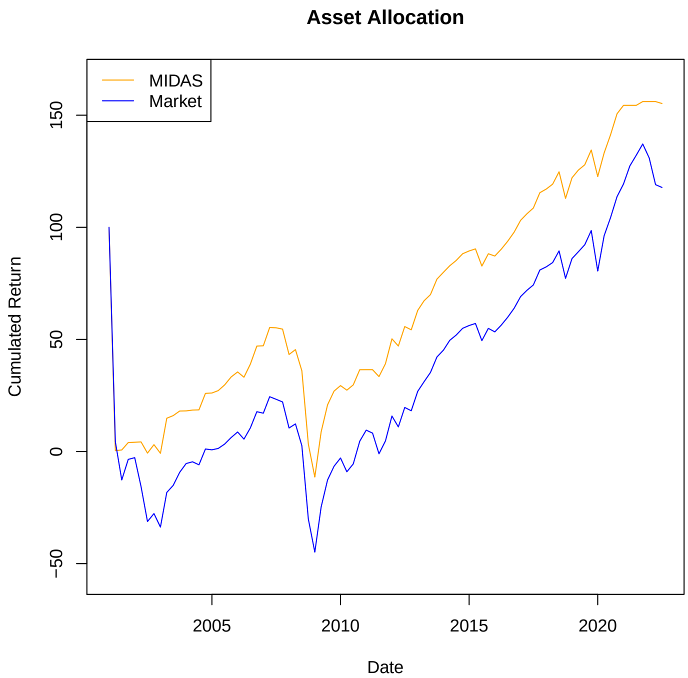


Figure 10: The picture depicts the cumulated quarterly returns for MIDAS asset allocation strategy (orange line) and the market portfolio (blue line).

References

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Armesto, M. T., Engemann, K. M., Owyang, M. T., et al. (2010). Forecasting with mixed frequencies. *Federal Reserve Bank of St. Louis Review*, 92(6):521–36.

Bordalo, P., Gennaioli, N., Kwon, S. Y., and Shleifer, A. (2021). Diagnostic bubbles. *Journal of Financial Economics*, 141(3):1060–1077.

Bordalo, P., Gennaioli, N., Porta, R. L., and Shleifer, A. (2019). Diagnostic expectations and stock returns. *The Journal of Finance*, 74(6):2839–2874.

Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics*, 138(1):291–311.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, X., Li, Y., Rapach, D. E., and Zhou, G. (2022). Anomalies and the expected market return. *The Journal of Finance*, 77(1):639–681.

Duffie, D. (2010). Presidential address: Asset price dynamics with slow-moving capital. *The Journal of finance*, 65(4):1237–1267.

Fama, E. F. and French, K. R. (1988). Permanent and temporary components of stock prices. *Journal of political Economy*, 96(2):246–273.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Favero, C. A., Gozluklu, A. E., and Tamoni, A. (2011). Demographic trends, the dividend-price ratio, and the predictability of long-run stock market returns. *Journal of Financial and Quantitative Analysis*, 46(5):1493–1520.

Favero, C. A., Melone, A., and Tamoni, A. (2019). Factor models with drifting prices. *Available at SSRN 3418352*.

Gennaioli, N. and Shleifer, A. (2018). A crisis of beliefs. In *A Crisis of Beliefs*. Princeton University Press.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The midas touch: Mixed data sampling regression models.

Goulding, C. L., Harvey, C. R., and Mazzoleni, M. G. (2023). Momentum turning points. *Journal of Financial Economics*, 149(3):378–406.

Hassan, T. A., Hollander, S., Van Lent, L., and Tahoun, A. (2019). Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics*, 134(4):2135–2202.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.

Newey, W. K. and West, K. D. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix.

Pesaran, M. H. and Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics*, 10(4):461–465.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2022). Measuring news sentiment. *Journal of econometrics*, 228(2):221–243.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.