Sparse Signals in Stock Returns and News

Abstract

We generate one-day-ahead return forecasts for the full universe of S&P 500 stocks using two sets of predictors: (i) lagged returns of all index stocks and (ii) estimates of news attention to business and finance topics. We follow Chinco, Clark-Joseph, and Ye 2019 and employ the Least Absolute Shrinkage and Selection Operator (LASSO). This approach allows us to handle a high-dimensional predictor set with relatively few observations by shrinking most coefficients to zero and retaining only a sparse subset of predictive signals. Our results show that (i) the LASSO forecast alone rarely delivers positive out-of-sample R^2 values, (ii) [INSERT RESULT OF COMBINED FORECAST], (iii) the number of selected predictors declines monotonically with the penalty parameter λ, and (iv) only about 8% of selected predictors exhibit time-series persistence, with substantial variation across the observation window.

1 Data

Stocks. We use all daily stock returns for firms in the S&P 500 over the period from December 28, 2015 to April 3, 2020. The data are obtained from CRSP using the PERMNO identifiers of the constituent firms. Returns $r_{i,t}$ are measured close-to-close.

News. Our text corpus consists of 69,612 news articles collected from the *New York Times*, *Reuters*, and *CNBC*. Articles are selected using finance-related keywords such as "stock market" and "financial market," to ensure relevance to asset pricing. The resulting documents are sent through a standard text processing pipeline. We estimate a Latent Dirichlet Allocation (LDA) model with $K = 20$ topics on this corpus. For each day t, we aggregate topic shares from articles to construct measures of daily topic attention. We extract AR(1) innovations from the daily topic share data. The resulting series captures unexpected shifts in attention to each topic.

Table 1: Summary statistics for returns and news

Variable	n	mean	sd	q05	median	q95
Return	520299	0.0004	0.0213	-0.0282	0.00087	0.0264
News	21380	0.00004	0.0239	-0.0305	-0.003973	0.0444

2 Econometric Setup

Forecast specification. We study whether lagged stock returns and topic-attention can forecast one-day-ahead returns. For firm i on day t, the forecasting equation is

$$r_{i,t+1} = \alpha_t + \mathbf{R}_{i,t}^{\top}\boldsymbol{\delta} + \mathbf{T}_{i,t}^{\top}\boldsymbol{\beta} + \varepsilon_{i,t+1}, \tag{1}$$

where $\mathbf{R}_{i,t}$ is a vector of lagged stock returns, $\mathbf{T}_{i,t}$ is a vector of lagged topic-attention, and $\varepsilon_{i,t+1}$ is the one-day-ahead forecast error.

Variable Selection. At each forecast date, the predictor set includes three daily lags of all S&P 500 stock returns and three lags of all $K = 20$ topic-attention variables. We would need to estimate 1561 parameters for each forecast. To estimate this using OLS we would need at least 1,561 observations per forecast date, which is impossible given that we only observe 1075 trading days. Even if we would have enough data, an OLS specification would not capture short-lived predictors but only those that are significant for longer. This motivates our use of LASSO regularization to perform automatic variable selection. In particular, we estimate the LASSO on a rolling window of $L = 30$ days. The optimization problem is then the following:

$$(\hat{\boldsymbol{\delta}}_t, \hat{\boldsymbol{\beta}}_t) = \arg\min_{\delta,\beta} \left\{ \frac{1}{L} \sum_{\tau=t-L+1}^{t-1} \sum_{i=1}^{N} \left(r_{i,\tau+1} - \alpha_t - \mathbf{R}_{i,\tau}^{\top}\boldsymbol{\delta} - \mathbf{T}_{i,\tau}^{\top}\boldsymbol{\beta}\right)^2 + \lambda\left(\|\boldsymbol{\delta}\|_1 + \|\boldsymbol{\beta}\|_1\right) \right\}. \tag{2}$$

We employ 10-fold cross-validation to choose the penalty parameter λ at each iteration. The set of variables selected by equation (2) is then used to compute the day-ahead return forecast for stock i using equation (1).

Specifically, our rolling window procedure works as follows: We estimate the LASSO using observations from days $t - 29$ through t, let the LASSO select the optimal subset of predictors for stock i, and generate a return forecast for stock i on day $t + 1$ using the selected predictors. We then advance the window by one day, re-estimate the model using days $t - 28$ through $t + 1$, and forecast returns for day $t + 2$. This process continues until we reach the end of our sample period. We perform this procedure for all stocks in our sample.

3 Results

Figure 1 plots the penalization parameter of each one-day ahead forecast against the number of variables selected by the LASSO. As expected, the number of predictors kept non-zero monotonically decreases in the size of λ.

Next, we compare the out-of-sample performance of the LASSO forecasts with a simple benchmark that always predicts the in-sample mean. Figure 2 shows that, on average, the LASSO forecasts underperform this historical mean benchmark.

In line with Chinco, Clark-Joseph, and Ye 2019, we continue by exploring whether augmenting a standard benchmark model with the predictors selected by the LASSO improves out-of-sample fit. We construct an AR(3) benchmark forecast for each stock and then expand it by including the LASSO-selected

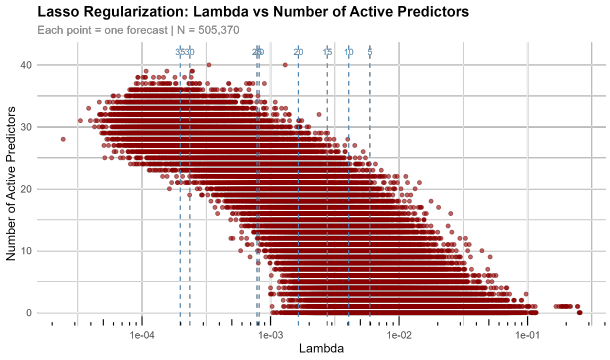

Figure 1: Number of predictors selected by LASSO as a function of λ.

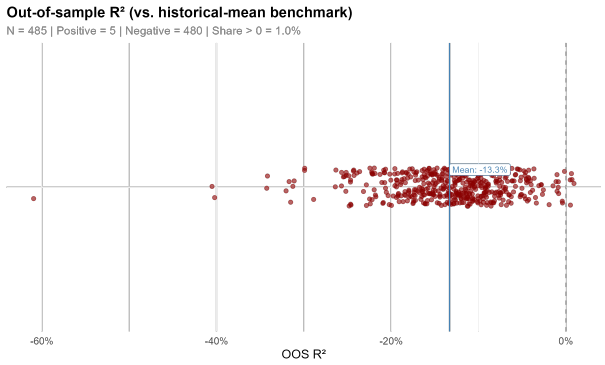

Figure 2: Out-of-sample R^2 of LASSO forecasts relative to the mean benchmark.

variables. If this combined model delivers a statistically significant increase in R^2_{oos}, we interpret it as evidence that the LASSO is capturing information not already contained in the benchmark model.

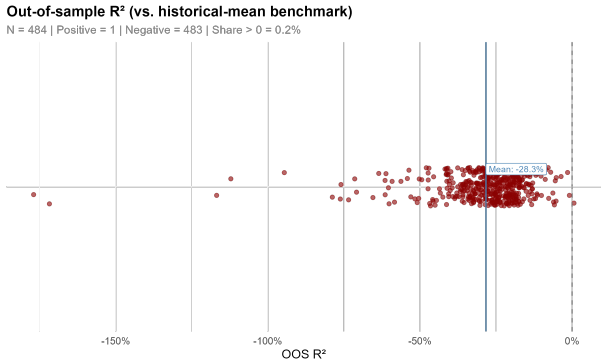

Figure 3: Out-of-sample R^2 of the AR(3) benchmark model.

Figure 3 shows that the benchmark model alone achieves, on average, a much lower R^2_{oos} than the LASSO specification. When the benchmark is augmented with the variables selected by LASSO, R^2_{oos} rises significantly. This indicates that our variable selection method captures information from the cross-section of stock returns and news attention that the AR(3) benchmark does not.

Finally, we analyze the time-series properties of the predictors selected by LASSO. Figure **??** shows the daily share of selected predictors that exhibit persistence. Persistence is assessed by estimating an AR(1) model for each selected predictor, using the same 30-day rolling window applied in the LASSO estimation. A predictor is classified as persistent if its autoregressive coefficient is statistically significant at the 5% level. On average, about 8% of predictors on a given day are persistent. This share varies considerably over time and tends to increase during periods of elevated market volatility (e.g., the COVID-19 crash).

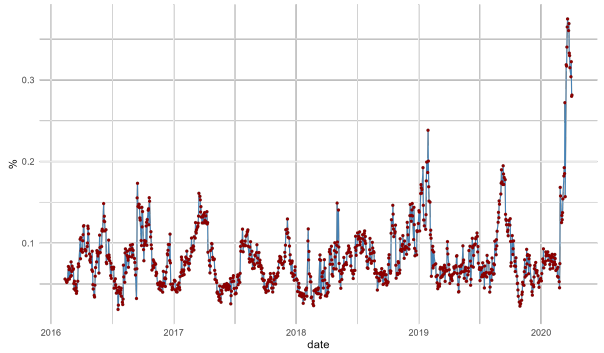

Figure 4: Daily share of LASSO-selected predictors classified as persistent.

References

Chinco, Alex, Adam D Clark-Joseph, and Mao Ye (2019). "Sparse signals in the cross-section of returns". In: *The Journal of Finance* 74.1, pp. 449–492.