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 number of selected predictors decreases monotonically with the penalty parameter λ, which tends to rise during volatile periods, (ii) combining the LASSO forecast with an AR(3) benchmark explains on average an additional 2.1% of daily return variation, and (iii) only about 7% of selected predictors exhibit time-series persistence, with higher persistence during episodes of market stress such as the COVID-19 crash.

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	484	0.0004	0.0213	-0.0282	0.00087	0.0264
News	20	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:

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

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 λ for each one-day-ahead forecast against the number of variables selected by the LASSO. As expected, the number of non-zero predictors decreases monotonically as λ increases.

Figure 2 reports the average daily value of λ. The penalty parameter varies substantially over time and tends to be higher during periods of elevated market volatility. This likely reflects the fact that, in such periods, the LASSO can exploit greater cross-sectional variation in returns and news topic attention.

Next, we assess the degree to which the LASSO forecasts capture the daily cross-sectional variation in individual stock returns. Following Chinco, Clark-Joseph, and Ye 2019, we estimate the stock-level regression model:

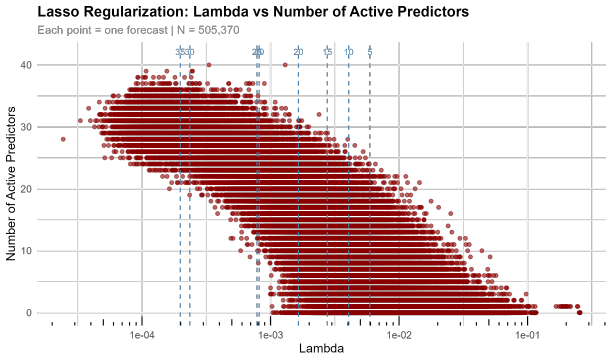

Figure 1: Number of predictors selected by LASSO as a function of λ.

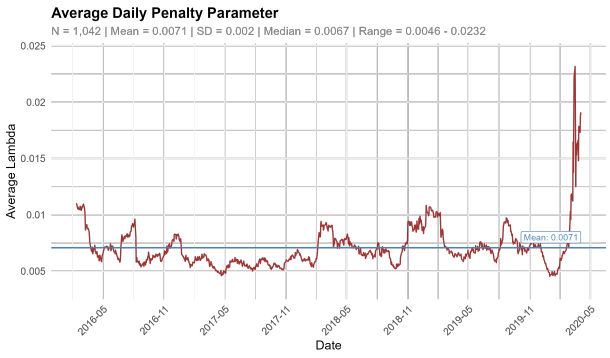

Figure 2: Average Daily Penalty Parameter λ.

$$r_{i,t} = \alpha_i + \beta_i z_{i,t} + \varepsilon_{i,t}, \tag{3}$$

$$z_{i,t} = \frac{f_{i,t} - \mu_i}{\sigma_i}, \tag{4}$$

where $r_{i,t}$ denotes the realized return of stock i on day t, and $f_{i,t}$ is the corresponding one-day-ahead forecast. Each forecast is standardized by subtracting its in-sample mean (μ_i) and dividing by its in-sample standard deviation (σ_i). This normalization ensures that regression coefficients are directly comparable across stocks and forecasting models.

Figures 3 and 4 present the R^2 values from equation (3), using $z_{i,t}$ constructed from the LASSO forecasts and, for comparison, from an autoregressive AR(3) benchmark estimated with the same rolling window procedure. On average, the LASSO forecasts explain 1.9% of the daily variation in returns, outperforming the AR(3) benchmark, which accounts for only 0.8%.

We next investigate the extent to which the information content of the LASSO and AR(3) forecasts overlaps. To this end, we re-estimate equation (3) including both predictors simultaneously. The results in Figure 5 show that the joint specification explains, on average, 2.8% of the daily variation in returns. Figure 6 reports the incremental change in R^2 at the stock level.

We next examine the time-series properties of the predictors selected by the LASSO. Figure 7 plots the daily fraction 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 employed in forecast construction:

$$r_{i,t} = \alpha_i + \rho_i r_{i,t-1} + u_{i,t}, \tag{5}$$

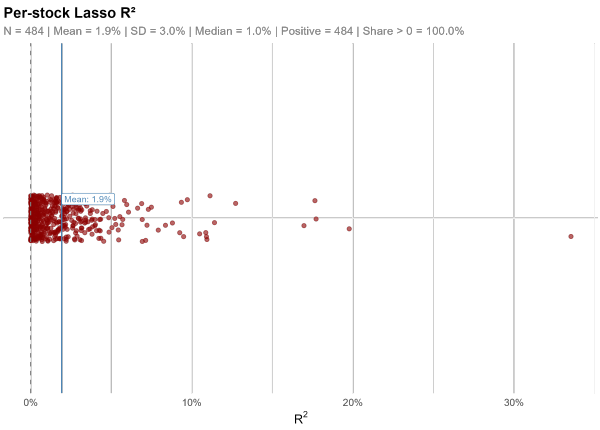

Figure 3: R^2 values from regressions using LASSO forecasts.

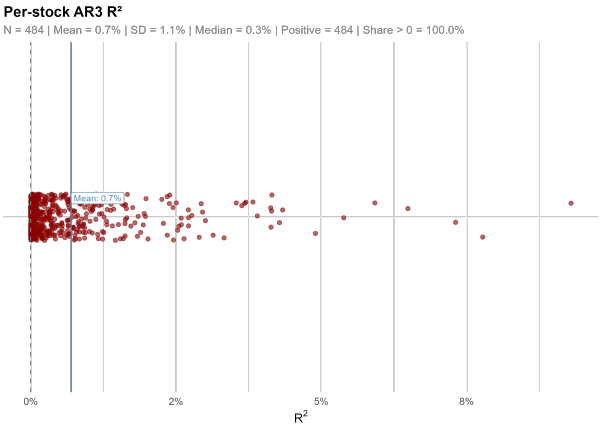

Figure 4: R^2 values from regressions using AR(3) benchmark forecasts.

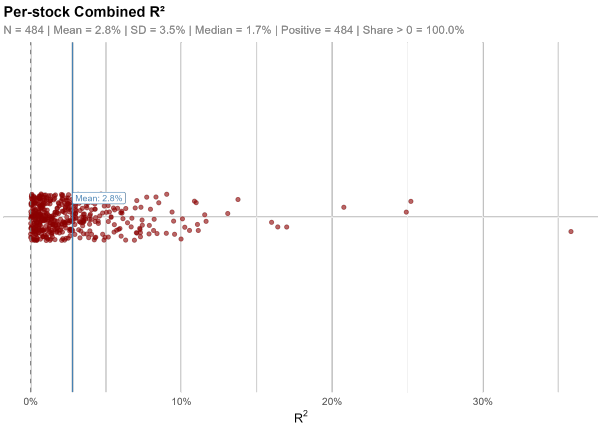

Figure 5: R^2 values from regressions using both LASSO and AR(3) forecasts.

where $r_{i,t}$ denotes the return of stock i on day t, ρ_i is the autoregressive coefficient, and $u_{i,t}$ is an error term. A predictor is classified as persistent if ρ_i is statistically significant at the 5% level. An identical specification is estimated for the news-based predictors, replacing returns with innovations in daily topic attention. The daily persistence share is defined as the ratio of predictors with significant ρ_i to the total number of predictors selected across all forecasts on that day. Because predictors may be included in multiple forecasts, the same variable can contribute more than once to the daily share. On average, 7.2% of predictors on a given day are persistent. This fraction fluctuates substantially over time and tends to increase

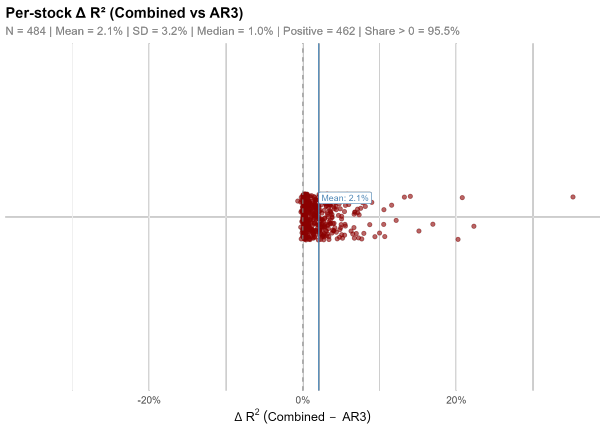

Figure 6: Change in R^2 from AR(3) only to AR(3) + LASSO.

during periods of heightened market volatility, such as the COVID-19 crash.

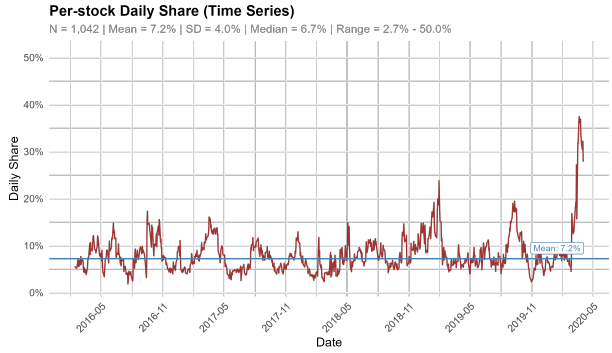

Figure 7: Daily share of Predictors selected by LASSO that are persistent.

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.