Forecasting and Managing Correlation Risks*

Tim Bollerslev†, Sophia Zhengzi Li‡, and Yushan Tang§

First Draft: June 2, 2022

This Version: April 15, 2023

Abstract

We propose a novel and easy-to-implement framework for forecasting correlation risks based on a large set of salient realized correlation features and the sparsity-encouraging LASSO technique. Considering the universe of S&P 500 stocks, we find that the new approach manifests in statistically superior out-of-sample forecasts compared to commonly used procedures. We further demonstrate how the forecasts translate into significant economic gains in the form of higher pairs trading profits, better equity premium predictions, more accurate portfolio risk targeting, and superior overall risk control and minimization.

JEL Classification: C13, C14, C52, C53, C55, C58.

Keywords: Correlation forecasting; high-frequency data; LASSO; risk targeting and control; pairs trading; equity premium prediction.

*We thank Clifton Green, Amit Goyal, Hao Jiang, Yuan Liao, Markus Pelger, Peixuan Yuan, Guofu Zhou, conference participants at the NBER Big Data and High-Performance Computing for Financial Economics Conference, the Shanghai Forum by Fudan University, and seminar participants at Durham University, Rutgers Business School, Renmin University, the Virtual Derivatives Workshop, and University of Rhode Island for their helpful comments and suggestions.

†Duke University, NBER and CREATES, Durham, NC 27708; E-mail: boller@duke.edu.

‡Rutgers Business School, Newark, NJ 07102; E-mail: zhengzi.li@business.rutgers.edu.

§Rutgers Business School, Newark, NJ 07102; E-mail: yushan.tang@rutgers.edu.

1. Introduction

A proper understanding of the likely co-movement among asset returns is crucial for asset pricing, portfolio construction, and risk management alike. We propose a new and easy-to-implement framework for forecasting correlation risks of stocks by explicitly focusing on out-of-sample prediction rather than in-sample statistical inference. Our approach succinctly combines so-called feature engineering and model fitting into a coherent framework. In the feature engineering step, we consider a number of variables that have previously been used in the literature, together with several new purposely designed features. In the model fitting step, we deliberately select the most useful features through the use of sparsity-encouraging estimation procedures. Our empirical results, based on intraday high-frequency data for a large sample of S&P 500 stocks spanning several decades, demonstrate highly statistically significant improvements in the accuracy of the out-of-sample forecasts compared to commonly used popular benchmark models. These statistical improvements also translate into substantial economical gains in portfolio construction, risk control, and return prediction, further underscoring the practical value of the new procedures.

Guided by the ideas of Kelly et al. (2022) among others, encouraging the use of *all* plausibly relevant predictors to improve the performance of machine learning-based procedures, we consider several alternative feature sets, each of which might ostensibly contain useful predictive information.[1] Consistent with well-documented autoregressive dynamic dependencies in correlations, the first feature set naturally consists of lagged daily, weekly, and monthly realized correlation measures, together with their semicorrelation extensions as formally defined below. The second feature set utilizes the common factor structures of stock returns by combining the high-frequency market data with low-frequency firm-characteristics through matrix projection. The third feature set considers a smoothed version of the realized correlations that explicitly incorporates sector information into the more traditional measures.

[1] See also Giglio et al. (2022) for a survey of recent methodological contributions in asset pricing involving the use of factor models and techniques adopted from machine learning.

Having defined the different feature sets, we next develop a correlation forecast model by simultaneously combining all features into a simple linear model. We consider both ordinary least squares (OLS) and the sparsity-encouraging least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) for fitting the models. We find that jointly fitting all features via OLS typically outperforms the popular benchmark HAR forecasting model (Corsi, 2009). However, sparsely fitting the features via LASSO results in additional significant improvements over all the OLS-based models. One key step of the LASSO fitting underlying these results consists of model selection via validation, in which we dynamically split the sample into separate training and validation sets used for estimation and prediction-model selection, respectively. This approach in turn reveals interesting time-varying patterns in the selected features in concert with varying economic and financial market conditions.

To more concretely evaluate the economic significance of the new correlation forecasting models, we also consider four distinct practical applications. In the first application, we study the augmented pairs trading strategy of Chen et al. (2019). We show that our new correlation forecasting models significantly improve the equal-weighted (value-weighted) strategy returns to 9.34% (8.85%) per annum from 3.63% (6.14%) based on HAR-model predictions. Our second application is motivated by Pollet and Wilson (2010), and the findings reported therein that the average correlation among individual stock returns is able to predict the aggregate return on the market. We show that the predicted correlations from our new models result in significantly stronger predictive power, with adjusted monthly R^2's for the equal-weighted (value-weighted) market index of 1.91% (2.12%), compared to 0.74% (0.60%) for the traditional HAR-based forecasts, and are comparable to some of the best market predictors. Our third application evaluates the risks of various long-short portfolio strategies and style-tilts, for which we compare the realized portfolio risks to their forecasted risks based on the competing risk models. We again find that our new model forecasts notably improve the consistency between realized and predicted risks, with the average risk-targeting ratio from LASSO across all strategies equal to 1.02, versus 0.83 from the HAR-based forecasts.

Our fourth, and final, application considers the construction of Global Minimum Variance (GMV) portfolios. We show that the new forecasting models achieve the lowest realized risks for the GMV portfolios. In addition, relying on the utility framework of Fleming et al. (2003), we show that a risk-averse beta-neutral GMV investor would be willing to sacrifice up to 6.07% annually to switch from the simple HAR-based forecasts to our new LASSO-based forecasts.

To address issues of robustness and sensitivity, we further split the full sample into smaller subsamples. Doing so, we find that although the overall quality of the model fits differs somewhat over different subsamples, the LASSO-based models consistently deliver superior out-of-sample forecasts, the forecasts for the recent COVID-19 period included. We also investigate possible gains afforded by the use of other economically motivated features based on a wide variety of different notions of firm linkages. However, none of these additional features significantly improves the accuracy of the out-of-sample forecasts compared to the forecasts based on our three specifically engineered feature sets. We also explore the use of alternative fitting procedures adopted from the machine learning (ML) literature, including Ridge Regression, Elastic Net, Principal Component Regression, and Feed-Forward Neural Networks. Again, none of these more complicated fitting procedures improves significantly on the simple-to-implement LASSO-based forecasts.

We make a number of contributions to the literature. First, our construction of the new feature sets specifically designed for correlation forecasting that succinctly combines the information from high-frequency market data and low-frequency fundamental data is decidedly new. Second, our use of sparsity-encouraging fitting techniques as a way to robustly exploit big data with many observations and features for more accurate large-scale correlation forecasting is similarly new to the literature. Third, our illustration of the strong economic gains afforded in a wide range of practical applications adds importantly to our understanding of the new procedures and further underscores the value of better risk forecasts more generally.

Our paper is related to several strands of literature. Most closely perhaps, to the large

financial econometrics literature, dating back to Engle (2002) and Tse and Tsui (2002), on modeling time-varying conditional correlations using GARCH-type models. Relatedly, the parametric models proposed by Cappiello et al. (2006) and Audrino and Trojani (2011) explicitly allow for asymmetric dynamic dependencies in the conditional correlations. These multivariate GARCH-type models have also subsequently been extended to incorporate realized variation measures in the construction of more accurate forecasts; see, e.g., Noureldin et al. (2012) along with the more recent paper by Bollerslev et al. (2020b) and the many other studies discussed therein. Instead, we focus directly on forecasting the realized correlation measures constructed from intraday data. The idea of directly modeling and forecasting realized variation measures dates back to Andersen et al. (2003), and it has been explored extensively in the literature since then, albeit mostly in the context of forecasting realized variances; see also the recent discussion and literature review in Bollerslev (2022). Meanwhile, our paper is distinct from this existing literature by allowing for more flexible dynamics and a much wider set of predictor variables than has hitherto been considered in the literature.

The paper also adds to the burgeoning literature on the use of ML techniques for financial decision makings. In addition to the already large existing literature devoted to return prediction, as exemplified by Rapach et al. (2013), Gu et al. (2020), Li and Rossi (2021), Bali et al. (2022), Chen et al. (2023), Kaniel et al. (2022) and Bali et al. (2023), our paper is most closely related to the recent literature on applying ML techniques for the purpose of risk management. The studies by Audrino and Knaus (2016), Bucci (2020), Li and Tang (2022), and Christensen et al. (2023), in particular, all rely on ML learning algorithms for univariate volatility forecasting, while Bollerslev et al. (2022a) do so for covariance matrix forecasting. Bollerslev et al. (2022b) similarly adapt ideas from ML to cluster stocks into groups based on their realized risk characteristics. Our new approach differs from these existing studies in two important dimensions. First, our work focuses on specifically designed and economically motivated new feature sets and a deliberately chosen fitting technique for building reliable forecasting models. Second, rather than focusing on pure statistical

assessments of the correlation predictions, we further demonstrate the economic value of the new procedures for a range of practical financial applications.

The paper is also related to the large finance literature on stronger comovements among certain types of stocks, including S&P 500 index constituents (Barberis et al., 2005), firms with similar institutional ownership (Pindyck and Rotemberg, 1993), firms with headquarters in the same geographical location (Pirinsky and Wang, 2006), and firms with similar analyst coverage (Muslu et al., 2014; Hameed et al., 2015; Israelsen, 2016), to name but a few. In contrast to all these studies, however, which primarily focus on causal relations between firm linkages and asset price movements, we focus explicitly on the prediction of future stock return correlations. In so doing, we show empirically that these previously established causal firm-connections provide limited predictive power over and above that afforded by the new specifically designed features proposed here.

The paper is organized as follows. Section 2 discusses the data and the different features underlying our empirical analyses. Section 3 summarizes the details of the machine learning methodology that we rely on. Section 4 compares the statistical out-of-sample forecasting performance of the new models to existing procedures. Section 5 highlights the economic significance of our new forecasting models by considering their uses in four distinct practical applications. Section 6 summarizes the results from additional robustness checks and empirical analyses. Section 7 concludes. Further details regarding various modeling choices along with additional summary statistics and data construction are deferred to an Appendix.

2. Data and variables

2.1. Data

We consider the universe of stocks that were ever constituents of the S&P 500 index and have full historical data from the NYSE Trade and Quote (TAQ) database between January 2000 and December 2020. We further require the stocks to be listed on the New York

Stock Exchange (NYSE), National Association of Securities Dealers Automated Quotations (NASDAQ), and the American Stock Exchange (AMEX) with share codes of 10 or 11, prices between $1 and $1,000, and daily number of trades greater than or equal to 100. To alleviate concerns about bid-ask bounce effects (Roll, 1984) that together with non-synchronous prices are well-known to induce downward biases in correlation estimation at ultra high frequencies (Epps, 1979), we rely on 15-minute intraday returns based on "coarsely" sampled mid-quote prices.[2] Some of our realized features require a history for their construction. To accommodate this, we use the data between 2000 and 2002 to compute the initial observations for these variables, fixing January 2003 as the common start date for our full-sample analysis. All in all, this leaves us with a final stock sample consisting of 417 unique S&P 500 stocks with full historical data for all features and response variables over the period from January 2003 to December 2020.

We also consider two additional sets of stock-level data for our feature construction and performance evaluation. The first set consists of 15 representative anomalies, including the 11 mispricing anomalies from Stambaugh et al. (2012), together with the traditional *CAPM Beta*, *Size*, *Book-to-Market*, and *Reversal* measures. We further complement this first set of anomalies with a set of 15 additional firm descriptors that have also been widely studied in the literature, including measures related to firm earnings, growth, and momentum. All of the anomalies are constructed following Green et al. (2013) and Chen and Zimmermann (2022) based on data from Compustat and CRSP. More detailed discussion and summary statistics are provided in the Appendix.

2.2. Response variable

Our main research objective centers on the development of better predictive models for monthly correlations. Accordingly, we rely on measures of monthly realized correlations as our response variable. Our use of the realized correlation measures may be formally

[2]As discussed further below, our choice of a 15-minute sampling frequency also mirrors that of many other recent studies; see, e.g., Fan et al. (2016) and Bollerslev et al. (2020a).

justified by the theory of quadratic variation. In particular, assuming that the joint dynamics of the stocks adhere to some underlying arbitrage-free Itô semimartingale vector process (Back, 2010), the true covariance matrix over a given time interval, like a day or a month, may then be consistently estimated by the summation of ever finer sampled within-interval cross-products of returns (Andersen et al., 2003; Barndorff-Nielsen and Shephard, 2004).

To set up the requisite notation, let $p_{i,\tau}$ denote the natural logarithm of stock i's price on day τ. Assume that intraday prices are observed at $n+1$ equally spaced time intervals from day τ to day $\tau+1$, say $p_{i,\tau},\ p_{i,\tau+1/n},\ ...,\ p_{i,\tau+1}$, with the corresponding returns denoted by $r_{i,\tau-1+k/n} \equiv p_{i,\tau-1+k/n} - p_{i,\tau-1+(k-1)/n}$ for $k = 1, ..., n$. The annualized daily realized variance for stock i on day τ and the realized covariance between stocks i and j on day τ are then simply constructed as:

$$RV_{i,\tau} \equiv 252 \times \sum_{k=1}^{n} r^2_{i,\tau-1+k/n}, \qquad RCov_{ij,\tau} \equiv 252 \times \sum_{k=1}^{n} r_{i,\tau-1+k/n} \cdot r_{j,\tau-1+k/n}. \tag{1}$$

Although the theory formally underlying consistency of the realized measures calls for the use of increasingly finer sample intraday returns, or $n \to \infty$, as previously noted to help mitigate the effects of non-synchronous trading and other market microstructure effects that might bias the estimates, in the results reported on below we deliberately rely on a "coarse" 15-minute sampling frequency, or $n = 27$.[3] In addition, to obtain an estimate of the variation for the entire day, we follow common practice in the literature (see, e.g., Hansen and Lunde, 2005) and include the overnight returns in the calculations of the full-day RV^d's and $RCov^d$'s. We obtain longer-run weekly and monthly measures, denoted by RV^w's, $RCov^w$'s, RV^m's and $RCov^m$'s in the sequel, by averaging the daily measures over one week and one month, respectively.

In line with the majority of the existing asset pricing literature, we will primarily focus on a one-month forecast horizon. Accordingly, to facilitate the notation we will often drop

[3]The volatility signature plot (as defined by Andersen et al., 2000) for the realized correlations in the Appendix further underscores the soundness of that choice.

the m superscript and use the time index t to refer to a month. Correspondingly, we will denote the monthly realized correlation matrix by RC_t, as implicitly defined by:[4]

$$RCov_t = \sqrt{RV_t} \cdot RC_t \cdot \sqrt{RV_t}, \tag{2}$$

where RV_t refers to the diagonal matrix with the month-t realized variances along the diagonal, and $RCov_t$ denotes the month-t realized covariance matrix. There is, of course, already a vast literature on modeling and forecasting realized variances; see, e.g., the recent discussions and literature reviews in Bollerslev et al. (2018a), Li and Tang (2022), and Christensen et al. (2023). We do not seek to add to this literature, instead focusing our empirical analyses exclusively on realized correlation forecasting and practical applications thereof.

Meanwhile, it is well-established that RV_t and RC_t tend to exhibit quite different dynamic dependencies (see, e.g., the discussion in Oh and Patton, 2016). To illustrate, Figure 1 plots the 12-month moving average of the cross-sectional means of the monthly realized correlations and realized variances over the full sample period. Even though the time series of realized correlations appear relatively stable, and clearly more so than the realized variances, the series still reveals non-trivial temporal variation in the cross-sectional average monthly realized correlations, with noticeable higher values in the aftermath of the 2008 financial crisis, as well as the recent COVID-19 period. Figure 2, which displays the unconditional distribution of all the monthly realized correlations, further highlights the dispersion in the individual monthly stock-pair correlations across time and stocks. We turn next to a discussion of the different features that we use in the construction of the new and improved forecasting models for said correlations.

[4]The CCC-GARCH model of Bollerslev (1990) and the DCC model of Engle (2002) also both rely on this same decomposition for forecasting the covariance matrix $RCov_t$ through the separate modeling and forecasting of RV_t and RC_t.

2.3. Features

We begin our analysis by constructing a series of input features that potentially contain predictive information for the one-month-ahead realized correlations. We consider three separate sets of features: ones inspired by HAR-type models, features constructed from common factors, and exponential-weighted realized features. We discuss each of the three specific feature sets in turn.

2.3.1. Realized variation and semivariation features

The Heterogeneous Autoregressive (HAR) model proposed by Corsi (2009) has arguably emerged as *the* benchmark model for realized volatility-based forecasting. The model is designed to succinctly capture the effects of heterogeneous short-, medium- and long-term signals with the forecasts of future volatilities based on linear combinations of lagged daily, weekly, and monthly realized volatilities. Motivated by the success of the traditional HAR model for forecasting RVs, our first feature set for forecasting the month $t+1$ correlation for stocks i and j similarly includes the lagged daily, weekly, and monthly realized correlations for the same two stocks, say $RC^{d}_{ij,t}$, $RC^{w}_{ij,t}$, and $RC^{m}_{ij,t}$. We will also refer to models based on just these three features as HAR models for short.

Moreover, motivated by Barndorff-Nielsen et al. (2010) and Patton and Sheppard (2015) and the empirical findings reported therein that HAR-type volatility forecasting models may be improved by including separate "bad," or downside, realized volatility measures constructed from the summation of squared negative high-frequency returns (see the recent discussion in Bollerslev, 2022), we also include daily, weekly and monthly negative realized semicorrelation measures in our first feature set. In particular, following Bollerslev et al. (2020a), the annualized daily realized negative semicovariance is naturally defined by restricting the summation in (1) to the products of the negative intraday returns only:

$$RCov^{d-}_{ij,\tau} = 252 \times \sum_{k=1}^{n} r_{i,\tau-1+k/n} 1_{\{r_{i,\tau-1+k/n}<0\}} r_{j,\tau-1+k/n} 1_{\{r_{j,\tau-1+k/n}<0\}}. \tag{3}$$

In parallel to the construction of the weekly and monthly realized covariances, the longer horizon $RCov^{-}$'s may similarly be obtained by averaging the daily $RCov^{-}$'s defined in (3) over the relevant horizon, in turn allowing for the construction of the corresponding daily, weekly, and monthly semicorrelation measures, say $RC^{d-}_{ij,t}$, $RC^{w-}_{ij,t}$, and $RC^{m-}_{ij,t}$. In total, this leaves us with six realized features, RC^{d}, RC^{w}, RC^{m}, RC^{d-}, RC^{w-}, and RC^{m-}, specifically designed to capture autoregressive dynamic dependencies as well as possible asymmetric responses to signed price movements. Following the nomenclature for realized volatility models in Patton and Sheppard (2015), we will refer to correlation forecasting models based on these six features as SHAR models in the sequel.

2.3.2. Factor-driven features

A slew of factor models have been proposed in the literature to account for the joint dependencies among returns through the decomposition of the total return co-movements into factor-driven and residual components. Notable examples include the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965), the three- and five-factor models of Fama and French (1993) and Fama and French (2015), the q-factor model of Hou et al. (2015), and the mispricing-factor model of Stambaugh and Yuan (2016). Several papers have further demonstrated the advantages of exploiting the factor structures in the formulation and estimation of covariance matrix forecasting models; see, e.g., Chan et al. (1999), Hansen et al. (2014), and Fan et al. (2016). Intuitively, if a factor model perfectly describes the common return variation, then the pairwise return co-movements should be entirely driven by the factors. In this situation, the factor-driven correlation components may naturally be interpreted as "de-noised" versions of the total correlations, and the inclusion of these components as additional features may therefore help in the prediction of the RC's more generally.

The true factors, of course, are not known. Instead, following Fama and French (2020) and Li et al. (2023) we rely on observable characteristics as factor loadings to "back out" a set

of factor-driven correlation features. Specifically, assuming that the $N \times 1$ vector of returns is driven by K common factors, the realized covariance matrix for the returns may then be decomposed as $RCov = L \cdot RCov^f \cdot L' + RCov^\epsilon$, where L denotes the $N \times K$ matrix of factor loadings, $RCov^f$ denotes the $K \times K$ factor covariance matrix, and $RCov^\epsilon$ refers to an $N \times N$ residual covariance matrix. The off-diagonal elements in $RCov^\epsilon$ account for the return covariation that is not explained by the common factors. Correspondingly, the factor-driven, or the "de-noised," covariance matrix may thus be expressed as $L \cdot RCov^f \cdot L' + Diag(RCov^\epsilon)$. The diagonal elements of that matrix naturally coincide with the diagonal elements of $RCov$. To obtain the off-diagonal elements, we substitute $(L'L)^{-1}L' \cdot RCov \cdot L(L'L)^{-1}$in place of $RCov^f$, resulting in the factor-driven covariance matrix estimate:

$$RCov^F = L(L'L)^{-1}L' \cdot RCov \cdot L(L'L)^{-1}L' + Diag(RCov^\epsilon). \tag{4}$$

This factor-driven covariance matrix at a given horizon may therefore be estimated from the observable characteristic matrix L and the realized covariance matrix $RCov$ at that same horizon.[5]

In the empirical results discussed below, we rely on a loading matrix L comprised of 15 representative anomalies and firm characteristics, including the 11 mispricing anomalies from Stambaugh et al. (2012) together with the usual *CAPM Beta*, *Size*, *Book-to-Market*, and *Reversal* measures.[6] Since the anomalies vary quite substantially in terms of their sample means and standard deviations, we follow Gu et al. (2020) and rank-transform each anomaly, except for the *CAPM Beta*, into the [-1, 1] interval. Having estimated the daily, weekly, and monthly $RCov^F$ covariance matrices, we finally covert each of them into correlation matrices. All said and done, this leaves us with three additional daily, weekly, and monthly factor-driven realized correlation features, denoted by FRC^d, FRC^w, and FRC^m, respectively.

[5] A similar approach has also previously been used by Fan et al. (2016) in their estimation of the factor covariance matrices from high-frequency factors formed by firm characteristic sorts.

[6] Table A.1 in the Appendix provides further details and summary statistics for each of the 15 anomaly variables.

We will refer to forecasting models based on the previous six realized features and these three additional factor-driven features as SHAR-F models.

2.3.3. Exponentially weighted realized features

Our third and final feature set is directly motivated by the Heterogeneous Exponential Realized Volatility (HExpRV) model recently proposed by Bollerslev et al. (2018a). This model provides a particularly simple framework for flexibly incorporating longer-run dynamic dependencies into realized-volatility-based forecasting by exploiting the use of exponentially weighted moving averages (EWMA) of lagged daily realized volatilities.

Analogously defined exponential-weighted negative semivariance, covariance, and negative semicovariance measures, may readily be constructed as:

$$\begin{aligned} ExpRV_{i,\tau}^{-\,CoM(\lambda)} &= \sum_{k=1}^{500} \frac{e^{-k\lambda}}{e^{-\lambda}+e^{-2\lambda}+...+e^{-500\lambda}} RV_{i,\tau-k+1}^{d-}, \\ ExpRCov_{ij,\tau}^{CoM(\lambda)} &= \sum_{k=1}^{500} \frac{e^{-k\lambda}}{e^{-\lambda}+e^{-2\lambda}+...+e^{-500\lambda}} RCov_{ij,\tau-k+1}^{d}, \\ ExpRCov_{ij,\tau}^{-\,CoM(\lambda)} &= \sum_{k=1}^{500} \frac{e^{-k\lambda}}{e^{-\lambda}+e^{-2\lambda}+...+e^{-500\lambda}} RCov_{ij,\tau-k+1}^{d-}, \end{aligned} \tag{5}$$

where λ defines the decay rate of the weights, and $CoM(\lambda)$ refers to the corresponding center-of-mass formally given by $CoM(\lambda) = e^{-\lambda}/(1-e^{-\lambda})$.[7] The center-of-mass effectively captures the "average" horizon of the lagged daily measures on which a given Exp measure is based. We construct $ExpRV$, $ExpRV^{-}$, $ExpRCov$, and $ExpRCov^{-}$ measures with center-of-mass equal to 1, 5, 21, and 63 trading days, corresponding to half-lifes of one day (d), one week (w), one month (m), and one quarter (q), respectively. We then convert the resulting variance and covariance terms defined by $ExpRV$ and $ExpRCov$, and $ExpRV^{-}$ and $ExpRCov^{-}$, respectively, into their respective exponential-weighted realized correlation and negative semicorrelation features, denoted by $ExpRC^{d}$, $ExpRC^{w}$, $ExpRC^{m}$, $ExpRC^{q}$, $ExpRC^{d-}$, $ExpRC^{w-}$, $ExpRC^{m-}$, and $ExpRC^{q-}$ in the sequel.

[7] Conversely, for a given center-of-mass, the corresponding decay rate equals $\lambda = log(1+1/CoM)$.

Additionally, in light of the strong within-sector asset co-movements previously documented in the literature (see, e.g., Fan et al., 2016; Herskovic et al., 2016), we also consider a series of sector-specific exponentially weighted realized features. We begin by dividing all of the stocks in our sample into sectors.[8] For each of the eight exponential features defined above and each sector Sc, we then compute the average value of the features across all of the stock pairs within a particular sector, denoting the resulting sector-specific exponentially weighted realized features by $ExpScRC^{d}$, $ExpScRC^{w}$, $ExpScRC^{m}$, $ExpScRC^{q}$, $ExpScRC^{d-}$, $ExpScRC^{w-}$, $ExpScRC^{m-}$, and $ExpScRC^{q-}$, respectively. If a pair of stocks belong to different sectors, their sector-specific features are all set to zero. This in turn leaves us with 16 additional features. We will denote SHAR-F models that include all of these additional features as SHAR-F-Exp models for short.

2.3.4. Feature summary statistics

The three different feature sets discussed above comprise a total of 25 unique correlation predictors.[9] Table 1 reports a series of descriptive statistics for each. There are several noticeable patterns. First, the sample means of all the features are positive, indicative of on-average positive stock co-movements. Second, the means of the factor-driven realized features (FRC's) are only slightly below those measured from the returns (RC's), suggesting that much of the co-movement among the stocks can indeed be accounted for by common factors. Third, the negative realized semi-features (RC^{-}'s, $ExpRC^{-}$'s, and $ExpScRC^{-}$'s) tend to have higher means than their unsigned counterparts (RC's, $ExpRC$'s, and $ExpScRC$'s), consistent with the idea of stronger asset co-movements during market downturns. Fourth, the exponential realized features ($ExpRC$'s) generally have lower standard deviations than the more traditional calendar-based realized features (RC's) due to the greater number of

[8]Our sector classifications are based on the first two digits of the stocks' Global Industry Classification Standard (GICS) codes from Compustat. The ten sectors include: Energy (10), Materials (15), Industrials (20), Consumer Discretionary (25), Consumer Staples (30), Health Care (35), Financials (40), Information Technology (45), Communication Services (50), and Utilities (55).

[9]To avoid any numerical issues, we further bound all of the empirically calculated features to lie in the unit interval, by replacing any values above (below) 1 (-1) with 1 (-1).

lagged daily realized correlations used in the smoothing. Lastly, the sector-specific exponential features ($ExpScRC$'s and $ExpScRC^{-}$'s) naturally have the lowest overall sample means, as they are fixed at zero for any stock pair belonging to different sectors.

3. Estimation methodology

Our competing forecasting models are based on linear specifications and two fitting procedures: Ordinary Least Squares (OLS) and the Least Absolute Shrinkage and Selection Operator (LASSO). The former entails a pre-defined set of predictors, whereas the latter performs variable selections to reduce the dimension of the feature sets as part of the model estimation.

3.1. Training and validation

In parallel to other machine learning algorithms, LASSO requires a validation set for tuning its hyperparameter(s). In the case of LASSO, this amounts to a single penalty parameter on the sum of the absolute slope coefficients, which succinctly controls the number of predictors used in the forecasting model. This hyperparameter should naturally be tuned based on the prediction error rather than the training error, to prevent LASSO from overfitting the training sample and performing poorly out-of-sample. Accordingly, we adopt a traditional training-validation-testing scheme for our hyperparameter calibration and model assessment.

To enhance the efficiency of the estimates, rather than fitting the models on a pair-by-pair basis, we instead fit "pooled models" based on the panel of all the pairwise realized correlations. Specifically, at the end of year t, we divide the sample into three parts: a training set consisting of data from year $t-4$ to year $t-1$, a validation set consisting of year t data, and a testing set consisting of year $t+1$ data. We then refit the models every year by rolling the training, validation, and testing sets one year forward. Given the $417 \times (417-1)/2 = 86,736$ unique stock pairs per month, each 4-year training sample thus includes 4,163,328 observations, allowing us to estimate models with many features. Importantly, our rolling-window scheme

also allows the features selected by LASSO to dynamically enter and exit the prediction models based on recent market conditions.

By contrast, our OLS-based prediction models rely on the same set of features throughout. Correspondingly, since the models estimated by OLS do not require validation sets for hyperparameter tuning, we use data from year $t-4$ to t for their estimation and the data in year $t+1$ for testing, thereby ensuring identical testing samples for the OLS and LASSO-based models.

3.2. Model fitting and LASSO

All of the forecasting models discussed in the main part of the paper are based on simple linear combinations of the different features, say $f(x_{ij,t};\theta) \equiv x'_{ij,t}\theta$, where $x'_{ij,t}$ denotes the feature vector for stock pair $(i,\, j)$ in month t and θ is the unknown parameter to be estimated. Unlike OLS, LASSO estimates θ through a penalized L_1 loss function. Specifically, to determine the best predictor for the monthly correlation for stock pair $(i,\, j)$, LASSO seeks to minimize:

$$\mathcal{L}^{LASSO}(\theta;\lambda) = \frac{1}{N}\sum_{(ij,t)\in\mathcal{T}}(RC^{m}_{ij,t+1} - x'_{ij,t}\theta)^2 + \lambda\sum_{p=1}^{P}|\theta_p|, \tag{6}$$

where $\mathcal{T}$ denotes a given training sample, N is the number of observations in that training sample, and λ refers to the shrinkage parameter that controls the degrees of penalty on the coefficients. For $\lambda = 0$, the LASSO estimator obviously collapses to standard OLS. However, when $\lambda > 0$, LASSO is capable of setting some of the coefficients to be exactly zero, thereby reducing the parameter estimation error and potentially also making the model easier to interpret.

Meanwhile, to allow for meaningful feature selection, we need to normalize the features to have comparable magnitudes, as otherwise a single λ would have vastly different shrinkage effects on different features, making the model impossible to tune. To prevent any look-ahead bias, we therefore further normalize each feature by its training-sample mean and standard deviation. Consistent with the rolling-window scheme detailed in the previous section, the

mean and standard deviation are calculated once per year. To allow for various levels of sparsity, we consider a wide range of different values of λ, choosing the final preferred model and values of θ and λ from this collection of models.

We turn next to a discussion of the resulting out-of-sample predictive performance obtained by the various OLS and LASSO-based models.

4. Out-of-sample forecast performance

We begin our assessment of the statistical out-of-sample forecasting performance, by demonstrating the forecast gains available by including additional features over-and-above the traditional HAR-type features in OLS-based models. We then illustrate the benefits of LASSO over the traditional OLS-based models. Finally, we analyze the sparsity of the LASSO-selected features and their dynamic patterns through time.

4.1. Performance evaluation measures

Following common practice in the literature, we focus our statistical assessment of the different models on their out-of-sample R^2's relative to the HAR model, which as previously noted has emerged as the benchmark model for realized volatility-based forecasting. Specifically:

$$R^{2,EW}_{OOS}(\theta) = 1 - \frac{\sum_{(ij,t)\in\mathcal{T}'}(RC^m_{ij,t} - \widehat{RC}^{m,\theta}_{ij,t})^2}{\sum_{(ij,t)\in\mathcal{T}'}(RC^m_{ij,t} - \widehat{RC}^{m,HAR}_{ij,t})^2}, \tag{7}$$

where $\widehat{RC}^{m,\theta}_{ij,t}$ refers to the forecasts from model θ, $\widehat{RC}^{m,HAR}_{ij,t}$ denotes the forecasts based on the traditional HAR features and OLS-based estimation, and $\mathcal{T}'$ defines the specific testing sample.[10] Accordingly, a positive $R^{2,EW}_{OOS}(\theta)$ indicates that model θ achieves smaller out-of-sample prediction mean squared errors than the benchmark HAR model. To ensure that our evaluation is not primarily driven by small-cap firms, we also calculate a value-weighted

[10]To ensure the predicted correlations are bounded between [-1, 1], we also apply an "insanity filter" and replace any predictions that fall outside that interval with 1 and -1, respectively.

version of the R^2_{OOS} as:

$$R^{2,VW}_{OOS}(\theta) = 1 - \frac{\sum_{(ij,t)\in\mathcal{T}'} \omega_{ij,t}(RC^m_{ij,t} - \widehat{RC}^{m,\theta}_{ij,t})^2}{\sum_{(ij,t)\in\mathcal{T}'} \omega_{ij,t}(RC^m_{ij,t} - \widehat{RC}^{m,HAR}_{ij,t})^2}, \tag{8}$$

where $\omega_{ij,t}$ denotes the product of the market capitalizations for stocks i and j normalized to sum to unity.[11]

To more formally assess the statistical significance of the numerical differences in the out-of-sample R^2s, we also implement a simple modified Diebold and Mariano (1995) (DM) test for pairwise comparison of two models based on the difference in the out-of-sample squared error losses. Specifically, for stock pair (i, j) in month t, we define the loss differential as $d_{ij,t} = (\hat{e}^{(1)}_{ij,t})^2 - (\hat{e}^{(2)}_{ij,t})^2$, where $\hat{e}^{(1)}_{ij,t}$ and $\hat{e}^{(2)}_{ij,t}$ denote the prediction errors from each of the two models. For each stock k, we then compute the average of the resulting loss differentials across all the stock pairs containing that stock. That is:

$$d_k = \frac{1}{N_k T} \sum_{ij,t} d_{ij,t} 1_{\{i=k \text{ or } j=k, i\neq j\}}, \tag{9}$$

where N_k denotes the number of stock pairs containing stock k in each month, and T refers to the total number of months in the entire testing sample. Our modified DM test statistic is obtained as $DM = \bar{d}/\hat{\sigma}_d$, where $\bar{d}$ and $\hat{\sigma}_d$ denote the cross-sectional mean and standard error of d_k. In parallel to the value-weighted R^2 defined above, we also construct a value-weighted version of this modified DM test statistic by applying the product-based weights to the loss differentials.

4.2. Forecast performance

Table 2 reports the resulting performance measures for the three increasingly more complex OLS-based models discussed in Section 2.3, together with the LASSO-based predictions

[11] The past two decades have witnessed the exceptional growth of many high-tech firms, and the 10 largest stocks in the S&P 500 now make up around 30% of the index's market value. Accordingly, we purposely winsorize market capitalization at 90% to control for the effect of mega firms.

outlined in Section 3.2. For ease of reference, the second column summarizes the specific features included in each of the models.

Looking first at Panel A and the out-of-sample R^2's, we observe the same ranking for the equal-weighted $R^{2,EW}_{OOS}$'s and the value-weighted $R^{2,VW}_{OOS}$'s defined in equations (7) and (8), respectively, suggesting that our results are not simply driven by extremely large or small firms. It is noteworthy that all of the models beat the traditional HAR model, albeit in the case of the SHAR model not by much. Adding the factor-driven realized features FRC's to the SHAR model results in more noticeable improvements. When we further expand the feature set to include the *ExpRC* and *ExpScRC* realized features, the resulting model outperforms all the other OLS-based models by quite wide margins, with the equal-weighted (value-weighted) R^2_{OOS} relative to the HAR equal to 9.82% (7.31%), highlighting the benefit of including the long-memory and within-sector spillover effects captured by these features.

Having established that the use of our new augmented feature sets can improve the out-of-sample performance through simple OLS-based fits, the row labeled LASSO demonstrates that additional improvements are available through the use of the LASSO algorithm. Intuitively, some features may not contribute to the forecasts over the entire sample. Accordingly, the dynamic regularization afforded by the LASSO algorithm may naturally help in the dynamic selection of the most useful features thereby avoiding overfitting. At the same time, even though LASSO does result in the highest R^2_{OOS} among all of the models, the improvements in the R^2_{OOS}'s compared to the OLS-based SHAR-F-Exp model that always includes all features may appear relatively minor.

Hence, to more formally assess whether these and the other improvements observed in Panel A are actually statistically significant, Panels B and C of Table 2 report the equal and value-weighted DM t-statistics detailed in Section 4.1 for the pairwise comparisons of the different models. A positive t-statistic in the table indicates that the "row model" outperforms the "column model." As the results show, all of the t-statistics for the LASSO-based predictions are positive and strongly statistically significant. In other words, these results

again corroborate that LASSO systematically produces the lowest forecast errors, and significantly so even compared to the all-encompassing SHAR-F-Exp model.

4.3. Feature selection

The sparsity engendered by the LASSO algorithm also brings with it easier interpretability of the forecasting models, in the form of the dynamic patterns of the selected features and their relative contribution to the overall predictions through time.

To illustrate, in each year in the testing sample, we calculate the absolute values of the estimated LASSO coefficients normalized by the sum of the absolute values across all stock pairs so that the scaled absolute coefficients add up to unity. Figure 3 presents the resulting relative contributions of the 25 realized features across years. The plot confirms that the LASSO models are indeed sparse-encouraging, with ten features selected per year on average. Among all realized features, $ExpRC^q$ is the most important predictor, being present in all testing samples and contributing around 50% overall to the predictions as measured by its scaled absolute coefficient. RC^w is also systematically selected, although it contributes less than around 5% on average. Among the three traditional HAR predictors, the lagged dependent variable RC^m is selected in 10 out of 13 years with an average importance of around 11%. The second set of frequently selected predictors includes the three short-term signals, FRC^d, FRC^w, and $ExpScRC^d$, each contributing around 4%. In other words, "denoising" the features by common factors and taking into account within-sector spillover effects are both valuable for correlation prediction. Lastly, although $ExpRC^m$ is only selected by LASSO in half of the testing years, its average contribution in those years is more than 15%, substantiating our early conjecture that certain features are only important under certain market conditions.

Looking across the columns in Figure 3, we also observe several interesting patterns in regard to the relative importance of different features through time. In particular, after training the model using data from the financial crisis period, LASSO selects the most

sparse feature set (4 out of 25 features) to form predictions for the year 2010. This directly illustrates the strength of LASSO over the OLS-based models which are not engineered to adapt to changing market conditions. Moreover, even though several long-term predictors are consistently selected by LASSO through time, different short-term signals enter and exit the models, in concert with the varying strength of the underlying signals. All said and done, however, the selected features do not change too dramatically from year to year, underscoring the overall stability of the LASSO-based approach.

5. Applications

The previous section demonstrates that the LASSO-based predictions result in significantly higher out-of-sample R^2's than all of the OLS-based models, the popular HAR model included. In this section, we further evaluate the economic significance of the predictions by considering four practical applications. In the first application, we construct an augmented pairs trading strategy following Chen et al. (2019), and show that the use of our new forecasting model substantially increases the return on the strategy. The second application builds on the finding of Pollet and Wilson (2010) that the average correlation among stocks can predict aggregate market returns. In the third application, we construct a series of long-short portfolios based on popular trading strategies and find that our model significantly improves the consistency between the realized and forecasted risks of the various portfolios. In the last application, we show that covariance matrix forecasts based on our new correlation forecasting models systematically achieve the lowest realized risks for the Global Minimum Variance (GMV) portfolio, resulting in substantial utility gains for a mean-variance investor.

5.1. Pairs trading

Pairs trading bets on the price convergence of pairs of stocks, or stocks with strong price co-movements. When price divergence occurs, pairs traders take a long position in the

underperforming stocks and a short position in the overperforming stocks, hoping to profit from the convergence. Building on Chen et al. (2019), we show how the better correlation forecasts obtained from our new model can help improve the performance of such strategies.

For each stock i in month t, we compute the historical correlation between that stock and each of the remaining 416 stocks in our S&P 500 universe using the time-series average monthly realized correlation between month $t-12$ and $t-1$. We define the top 20 stocks with the highest historical correlation with stock i as its pairs. We then compute the historical correlation between stock i and its pair portfolio, say $RC^h_{i,t}$, as the equal-weighted average of the historical correlations between stock i and each of its pairs.

Denoting the equal-weighted average return across the stocks in the pair portfolio by $PRet_{i,t}$, mimicking Chen et al. (2019), we measure return divergence $RetDiff_{i,t}$ between stock i and its pair portfolio by:

$$RetDiff_{i,t} = \beta_{i,t}(PRet_{i,t} - r_{f,t}) - (Ret_{i,t} - r_{f,t}), \tag{10}$$

where $r_{f,t}$ denotes the risk-free rate, $Ret_{i,t}$ denotes return on stock i, and $\beta_{i,t}$ is the regression coefficient from regressing stock i's returns on its pair portfolio returns using daily data between month $t-12$ and $t-1$. The intuition behind the pairs trading strategy is simple: if in month t stock i's return $Ret_{i,t}$ is above (below) its pair portfolio return $PRet_{i,t}$ after risk adjustment, the price of the stock is likely overvalued (undervalued) and we expect it to go down (up) next month. To implement the strategy, by the end of each month, we sort stocks into quintile portfolios by *RetDiff* and form a long-short portfolio by buying stocks with high *RetDiff* and selling stocks with low *RetDiff*.

A key implicit assumption behind the above pairs trading strategy is the persistence of correlations; i.e., a pair portfolio identified using the historical correlations for stock i is expected to comove strongly with that same stock in month $t+1$, even if it exhibits temporary price divergence in month t. This assumption may not always hold true, especially during periods with volatile market conditions. Accordingly, we aim to improve the strategy by

explicitly incorporating correlation predictions into the portfolio construction. Specifically, for stock i in month t, we compute its predicted correlation with its pair portfolio in month $t+1$ as the equal-weighted average predicted correlation with its pair stocks. To keep the presentation manageable, we focus on the predictions from the HAR and LASSO-based models, but similar results are available for the other forecasting models discussed above. The difference between the predicted and historical correlations, $\Delta RC^{\theta}_{i,t} = \widehat{RC}^{\theta}_{i,t+1} - RC^{h}_{i,t}$, where $\widehat{RC}^{\theta}_{i,t+1}$ refers to the prediction from model θ available by the end of month t, naturally captures the persistence of correlation between stock i and its pair portfolio. To identify stocks whose prices are more likely to converge to their historical pairs next month, we sort all the stocks into quintiles by their ΔRC^{θ} in each month, keeping only the subset of stocks in the highest quintile. We then sort these stocks into five quintiles by *RetDiff* to form a long-short portfolio.

Panels A and B of Table 3 report the annualized equal-weighted and value-weighted average returns of these quintile-sorted portfolios, along with the returns and t-statistics of the spread portfolios.[12] Looking first at the "Unconditional" strategy based on all of the S&P 500 stocks in our sample that does not take into account the persistence of the correlations, we see that it results in rather poorly performing portfolios, with an equal-weighted (value-weighted) spread of 1.15% (-1.20%) per year and a t-statistic of 0.47 (-0.45).[13] By contrast, the long-short portfolios formed by the subset of stocks within the highest ΔRC^{θ} quintile based on the HAR and LASSO correlation predictions all result in average positive returns. Intuitively, if model θ is good at predicting next month's correlations, the return convergence for the resulting pair stocks is likely to be stronger. For stocks within the highest ΔRC quintile based on HAR, the equal-weighted (value-weighted) return spread sorted by *RetDiff* does indeed increase to 3.63% (6.14%) per year, albeit statistically insignificant. Meanwhile, for the stocks within the

[12]For value-weighted quintile portfolios, we further impose the restriction that none of the weights exceed 10% to ensure that the portfolio is diversified.

[13]The pairs trading strategy in Chen et al. (2019) is formed based on a different stock universe over the extended period from January 1931 to December 2007. However, from Figure 1 in their paper, the value-weighted strategy also performed poorly towards the end of their sample period, with negative annual returns in six out of nine years between 1999 and 2007. Their equal-weighted strategy performed better than the value-weighted counterpart, yet it still generated a negative return in 2007 at the end of their sample.

highest ΔRC quintile based on LASSO, the equal-weighted (value-weighted) return spread increases to an impressive 9.34% (8.85%) per year, with a t-statistic of 2.30 (2.20). This much stronger pairs trading profit among high-ΔRC^{LASSO} stocks thus again corroborates the superior predictive power of our LASSO-based correlation forecasting model and illuminates the value of incorporating better correlation predictions into pairs trading.

To further underscore the significance of the results, we also conduct a series of predictive Fama-MacBeth cross-sectional regressions, controlling for the four Fama-French-Carhart risk factors.[14] To allow for a direct comparison with the previous single-sorting results, we transform *RetDiff* into discrete values of [1, 2, 3, 4, 5], corresponding to the five stock quintiles sort. Panel C of Table 3 reports the resulting estimated regression coefficients and t-statistics under several model specifications. Consistent with our previous portfolio-based analyses, *RetDiff* exhibits the most significant predictive power among stocks with the highest price convergence rate (ΔRC) as predicted by LASSO. For example, take the regressions with control variables reported in the even-numbered columns. The coefficient for *RetDiff* increases from 0.04 (t-statistic 0.97) in column (2) for the full stock sample, to 0.12 (t-statistic 1.80) in column (4) for stocks within the highest ΔRC^{HAR} qunitile, to 0.16 (t-statistic 2.33) in column (6) for stocks within the highest ΔRC^{LASSO} quintile. Given *RetDiff* takes integer values between 1 and 5 representing the five *RetDiff* quintiles, moving up one quintile is associated with an increase in annual return of 1.92% (0.16×12).

To illustrate how the long-short portfolios based on *RetDiff* perform over time, we compute the cumulative profits of the equal-weighted and value-weighted strategies with an initial investment of $W_1 = \$1$. Specifically:

$$W_{t+1} = W_t \times (1 + R^H_{t+1} - R^L_{t+1} + r_{f,t+1}), \tag{11}$$

[14] We estimate *Beta* by regressing monthly stock excess returns on monthly market excess returns using a 60-month rolling window. *Size* is the natural logarithm of the market value of equity, estimated by the product of the closing price and the number of shares outstanding. *Book-to-market ratio* is the ratio of the book value of common equity to the market value of equity. *Momentum* is estimated by the cumulative return over the past 2 to 12 months.

where R^H_{t+1} and R^L_{t+1} are the monthly returns on the fifth and first quintile portfolios, respectively, for a given strategy. Panels A and B of Figure 4 plot the resulting trajectories of W_{t+1} starting from February 2008. The portfolio value based on stocks within the highest ΔRC^{LASSO} quintile continues to rise throughout the entire sample without major drawdowns. In contrast, the value of strategies formed by stocks within the highest ΔRC^{HAR} quintile increases very minimally, and those of the unconditional strategies remain almost flat over time. Altogether, the results lend further support to integrating more accurate correlation forecasts from our LASSO-based model for enhanced pairs trading performance.

5.2. Equity premium predictions

Pollet and Wilson (2010) argue that the average correlation among stocks, say $AvgCorr$, manifests aggregate systematic risks and therefore should help predict future aggregate market returns.[15] In their empirical analyses, they estimate the expected future average correlation $E_t(AvgCorr_{t+1})$ based on the pairwise correlations computed from lagged daily returns. By that same logic, the use of a superior correlation prediction model for $E_t(AvgCorr_{t+1})$ should naturally result in stronger predictive power for the equity premium.

To test this conjecture, we estimate equal-weighted and value-weighted $E_t(AvgCorr_{t+1})$ from predicted monthly realized correlation for each stock pair as:

$$\begin{aligned} AvgCorr_t^{\theta,EW} &= \sum_{i=1}^{N} \sum_{j=1,\, j\neq i}^{N} \frac{1}{N(N-1)} \widehat{RC}_{ij,t+1}^{m,\theta}, \\ AvgCorr_t^{\theta,VW} &= \sum_{i=1}^{N} \sum_{j=1,\, j\neq i}^{N} \omega_{ij,t} \widehat{RC}_{ij,t+1}^{m,\theta}, \end{aligned} \tag{12}$$

where $\widehat{RC}_{ij,t+1}^{m,\theta}$ is the predicted realized correlation for month $t+1$ from model θ using information up to month t, EW and VW denote the equal and value-weighted average correlations respectively, N is the total number of stocks, and $\omega_{ij,t}$ is the product of the

[15] For a survey of the latest literature on market return predictability, see, e.g., Rapach and Zhou (2022).

market capitalizations for stocks i and j normalized to sum to unity in each month.[16] We again focus on the HAR and LASSO-based predictions. As a benchmark, we also construct historical equal and value-weighted average correlation based on the realized correlations RC in month t.

Table 4 reports the results from predictive regressions for the one-month-ahead excess return on the CRSP value-weighted market index. In addition to the different equal- or value-weighted average correlation measures, we also report results from multiple regressions in which we include the eight commonly used macroeconomic predictors from Welch and Goyal (2008) as controls.[17] From the simple regressions in the first three columns of each panel, the $AvgCorr^{H,EW}$ and $AvgCorr^{H,VW}$ constructed from the historical realized correlation are never significant. The $AvgCorr$ measures constructed from the HAR-model predictions exhibit stronger predictive power, albeit still insignificant with t-statistic of 1.47 (1.39) and adjusted R^2 of 0.74% (0.60%) for the equal-weighted (value-weighted) measure in the simple regression without controls. By contrast, the $AvgCorr$ measures constructed from the LASSO-based correlation predictions all provide significant information for predicting future market returns with t-statistics around two and much higher adjusted R^2's at 1.91% and 2.12%, respectively. Importantly, these same qualitative results for the simple regressions carry over to the multiple regressions with controls. In particular, only the $AvgCorr^{LASSO}$ measures positively and significantly predict next month's market excess return, with t-statistics of 2.40 and 2.66 for the equal and value-weighted measures, respectively. Given the average cross-sectional standard deviation of $AvgCorr^{LASSO,EW}$ ($AvgCorr^{LASSO,VW}$) is 0.0596 (0.0609), this means that a one-standard-deviation increase in the average correlation predicts a nontrivial rise

[16] We further winsorized the firm capitalization at the 90% level for each month to avoid weighting excessively on mega firms.

[17] All macroeconomic predictors are downloaded from Amit Goyal's personal website. *dp* is the dividend-price ratio computed as the difference between the log of 12-month moving sums of dividends and the log of prices of S&P 500 index. *ep* is the earnings-price ratio calculated as the difference between the log of 12-month moving sums of earnings and the log of prices of the S&P 500 index. *bm* is the book-to-market ratio defined as the ratio of book value to market value for the Dow Jones Industrial Average. The net equity expansion *ntis* is a ratio of 12-month moving sums of net issues by NYSE listed stocks divided by the total end-of-year market capitalization of NYSE stocks. *tbl* is the 3-month Treasury Bill rate. *tms* is the difference between the long-term yield on government bonds and the Treasury bill. *dfy* is the difference between BAA- and AAA-rated corporate bond yields. *svar* is the sum of squared daily returns on the S&P 500.

in the market excess return of $0.0596 \times 0.24 \times 12 = 17.2\%$ ($0.0609 \times 0.25 \times 12 = 18.3\%$) per annum. As such, these results again highlight the economic gains afforded by the more accurate LASSO-based *RC* forecasting model.

5.3. Risk targeting

Having access to more accurate correlation predictions allows portfolio managers to better assess the risk of their strategies while maximizing returns to investors. To illustrate, we consider a portfolio manager who allocates her funds into N risky assets based on a long-short trading strategy. At the beginning of month t, she sets the portfolio weight for stock i to $\omega_{i,t} = 1$ (-1) if the stock is in the long-leg (short-leg) of the strategy. To assess the risk of her portfolio ex-ante, she relies on the forecasts for the monthly covariance matrix based on the predicted correlations and volatilities.

Specifically, relying on the decomposition in (2), we predict the covariance matrix for month t based on:

$$\widehat{RCov}_t^{\theta} = \sqrt{\widehat{RV}_t} \cdot \widehat{RC}_t^{\theta} \cdot \sqrt{\widehat{RV}_t}, \tag{13}$$

where $\widehat{RV}_t$ denotes the diagonal matrix of predicted realized variances, and $\widehat{RC}_t^{\theta}$ denotes the predicted correlation matrix from model θ. To focus on the correlation forecasts, we purposely use the popular HAR-based forecasting model to predict the monthly realized volatilities regardless of the choice of correlation forecasting models. More sophisticated volatility forecasting models, including machine learning-based procedures (e.g., Li and Tang, 2022), have obviously been proposed in the literature. However, as previously noted, the simple HAR model has proven quite effective, and it has emerged as *the* benchmark model in the realized volatility forecasting literature. Accordingly, we purposely do not seek to optimize the model for $\widehat{RV}_t$, instead focusing on the comparison of different correlation forecasts $\widehat{RC}_t^{\theta}$, where we again restrict our main discussion to $\theta =$ HAR or LASSO.

Utilizing $\widehat{RCov}_t^{\theta}$, the expected risk of the portfolio ω_t under model θ is simply given by

$\omega_t' \widehat{RCov}_t^{\theta} \omega_t$. After the realization of the portfolio's return in month t, the portfolio manager can observe the ex-post portfolio risk and compare it with her expectation, in the form of the risk-targeting ratio: $\omega_t' \widehat{RCov}_t^{\theta} \omega_t / \omega_t' RCov_t \omega_t$. Averaging these risk-targeting ratios across different testing samples, we obtain the summary measure:

$$AvgRatio^{\theta} = \frac{1}{T} \sum_{t=1}^{T} \frac{\omega_t' \widehat{RCov}_t^{\theta} \omega_t}{\omega_t' RCov_t \omega_t}. \tag{14}$$

Since the portfolio weights are pre-determined and the volatility forecasts are generated from the same RV forecasting model, the more accurate the correlation forecasts, the closer this average risk targeting ratio should be to unity.

We proceed to compare the $AvgRatio^{\theta}$ ratios defined in (14) based on the HAR and LASSO correlation forecasting models by considering 15 different strategies formed by the anomalies previously used in our feature construction in Section 2.3.2. As the results in Figure 5 clearly show, the risk-targeting ratios for the LASSO-based forecasts are generally much closer to unity than the ones for the HAR-based forecasts.[18] Specifically, while the overall mean of the average risk-targeting ratios across all of the strategies equals 1.02 for LASSO, it equals 0.83 for HAR, indicative of systematic underestimation of the true risks. Further along these lines, the target of unity is included in the bootstrapped 95% confidence intervals for six strategies for LASSO, whereas for only one strategy is unity included in the 95% confidence intervals for the HAR-based forecasts.[19] These differences thus yet again underscore the superiority of the LASSO-based forecasts in the context of portfolio risk evaluation and targeting.

[18] We also evaluate the risk-targeting performance based on 15 additional anomalies pertaining to various aspects of firms' operation and market performance, as further detailed in Table A.2. The ratios for these additional strategies, displayed in Figure A.2 in the Appendix, exhibit very similar patterns.

[19] For each strategy, we have 155-month observations of the risk-targeting ratio. We perform 1,000 iterations of the bootstrap, where for each iteration we draw 155 ratios from the original data with replacement. For each bootstrapped sample, we then compute the average risk-target ratio, using the 2.5th and 97.5th percentiles of the resulting 1,000 average risk-targeted ratios as our bootstrapped confidence interval.

5.4. Global minimum variance portfolio

In our final application, we consider estimates of the Global Minimum Variance (GMV) portfolio. The GMV is often used for evaluating covariance matrix forecasts, as the portfolio weights only depend on the covariance matrix and not on the expected return, thus allowing for a "clean" comparison.[20]

Specifically, consider a risk-averse investor who allocates her wealth into N stocks with the goal of minimizing the total risk of her portfolio based on the covariance matrix forecast $\widehat{RCov}_t^{\theta}$ from some forecasting model θ. By standard arguments, her optimal portfolio weight vector may readily be constructed as:

$$\omega_t^{\theta} = \frac{(\widehat{RCov}_t^{\theta})^{-1}}{\mathbf{1}'(\widehat{RCov}_t^{\theta})^{-1}\mathbf{1}}, \tag{15}$$

where $\mathbf{1}$ denotes an $N \times 1$ vector of ones.[21] Calculating these optimal weights for each month t in the testing samples, we construct the resulting monthly portfolio returns $\omega_t^{\theta\prime} r_t$, realized portfolio risks $\sqrt{\omega_t^{\theta\prime} RCov_t \omega_t^{\theta}}$, and portfolio Sharpe ratios $(\omega_t^{\theta\prime} r_t - r_{f,t})/\sqrt{\omega_t^{\theta\prime} RCov_t \omega_t^{\theta}}$. To keep the presentation manageable, we again focus on the HAR and LASSO-based correlation forecasting models, together with the traditional HAR model for forecasting the volatilities, in line with the analysis in the previous Section 5.3.

To further assess the economic gains, we follow the framework of Fleming et al. (2003), as more recently extended by Bollerslev et al. (2018b), and assume that the investor's realized utility for the month-t return provided by model θ may be expressed as:

$$U(r_t^{\theta}, \gamma) = (1 + r_t^{\theta}) - \frac{\gamma}{2(1+\gamma)}(1 + r_t^{\theta})^2, \tag{16}$$

where γ refers to her level of risk aversion. Accordingly, the economic value of switching from

[20]Interestingly, Jagannathan and Ma (2003) also find that GMV portfolios often achieve higher out-of-sample Sharpe ratios than mean-variance optimized tangent portfolios.

[21]This optimal solution requires that the covariance matrix estimate is positive definite. For the LASSO-based correlation matrix forecasts that are non-positive definite, we apply a simple convexity correction as detailed in the Appendix.

forecasting model θ_1 to model θ_2 may naturally be measured by solving for Δ_γ in:

$$\sum_{t=1}^{T} U(r_t^{\theta_1}, \gamma) = \sum_{t=1}^{T} U(r_t^{\theta_2} - \Delta_\gamma, \gamma), \tag{17}$$

with Δ_γ interpretable as the return the investor would be willing to give up for using the better correlation forecasting model.

The summary table included in Figure 6 presents the resulting average realized portfolio returns, standard deviations, and Sharpe ratios, together with the estimates of Δ_γ.[22] Even though the return on the portfolio never enters the optimization problem, the average realized monthly return for the LASSO-based portfolios turns out to be slightly higher than for the HAR portfolios. This, of course, is merely by "luck." Importantly, however, LASSO-based portfolios also result in lower average risk than HAR portfolios, with an average portfolio standard deviation of 34.49% versus 36.42%. To more directly illustrate this, Figure 6 provides a scatter plot of the monthly portfolio standard deviations (annualized) for the HAR and LASSO-based GMV portfolios. As the figure shows, most of the data points are above the 45-degree line, indicating that the LASSO-based GMV portfolios typically result in lower realized risks than the HAR-based GMV portfolios.[23] The lower risk for the LASSO-based portfolios, together with the slightly higher average returns, naturally result in a higher average Sharpe ratio of 0.48, compared to 0.36 for the HAR-based portfolios. This also translates into nontrivial economic gains. In particular, from the last three columns in the table, an investor with risk aversion $\gamma = 2$ would be willing to give up 0.77% return per annum for access to our LASSO-based forecasts versus relying on the HAR-based forecasts. This economic gain further increases to 0.98% (1.37%) for a more risk-averse investor with $\gamma = 5$ ($\gamma = 10$).

[22]It is worth noting that even though we do not impose any constraint on the weights (as done in, e.g., Chan et al., 1999), neither of the two models generates especially extreme weights, as evidenced by the minimum portfolio weights for both the HAR and LASSO-based predictions never falling below -10%, and the maximum portfolio weights being around 30% for both models across all stocks and months in the sample.

[23]Concretely, the standard deviations of the HAR-based GMV portfolios exceed those of the LASSO-based GMV portfolios for 77.42% of the months in the sample.

Motivated by a prototypical market-neutral strategy in which managers want to neutralize market risk, we also consider a beta-neutral GMV analysis following Cosemans et al. (2016). In this setup, we augment the traditional GMV optimization problem with the additional constraint that the portfolio's beta equals zero. That is:

$$\frac{\omega_t' \widehat{RCov}_t^{\theta} m_t}{m_t' \widehat{RCov}_t^{\theta} m_t} = 0, \tag{18}$$

where m_t denotes the $N \times 1$ vector of firm market capitalization normalized to sum to unity. The returns, risks, and Sharpe ratios of the resulting beta-neutral GMV portfolios, reported in the table included in Figure 7, mirror the patterns for the original GMV portfolios. In particular, the LASSO-based portfolios again attain the lowest average risk and highest average Sharpe ratio. Importantly, the average of the actually realized monthly betas from the LASSO-based predictions, calculated as $\omega_t' RCov_t m_t / m_t' RCov_t m_t$, equals 0.05 with a robust t-statistic of 1.58, suggesting that the portfolios are indeed market neutral. By contrast, the average realized beta for the HAR-based predictions equals -0.20 with a robust t-statistic of -7.19, indicating significant negative exposure to market risk. To further illustrate these differences, Figure 7 displays time series plots of the monthly realized betas. As the figure shows, the realized betas for LASSO are seemingly symmetric around zero, while the betas for HAR appear systematically below zero. Interestingly, the imposition of the beta-neutral constraint also substantially increases the utility gains to 1.81% (6.07%) per annum for an investor with risk aversion $\gamma = 2$ ($\gamma = 10$). In other words, not only do the LASSO-based correlation forecasts allow a portfolio manager to better attain her risk target, the forecasts also allow her to better control her exposure to market risk.

6. Robustness

6.1. Subsample analysis

The empirical analyses discussed above relied on the full testing sample for comparing and contrasting the forecasting models. To illustrate that the superior performance remains intact across different sub-samples, we divide the full testing sample into three subperiods: 2008-2011, 2012-2015, and 2016-2020.

The resulting out-of-sample performance for each of the three periods is summarized in Table 5. Looking first at Panel A and the equal-weighted R^2_{OOS}'s, we see that during the first subperiod between 2008 and 2011, which covers the financial crisis and its aftermath, LASSO again achieves the highest relative R^2_{OOS} of 7.87%, while the OLS-based model with all the 25 main features included has the second highest relative R^2_{OOS} of 6.95%. Put differently, the sparsity engendered by LASSO improves the accuracy of the out-of-sample predictions when the market is volatile by focusing on the predictors that matter the most. For the second subperiod between 2012 and 2015, LASSO remains the best model with a relative R^2_{OOS} of 10.70%. For the most recent period spanning 2016 to 2020, although the R^2_{OOS} for LASSO is slightly below the OLS-based model that involves all the features, it clearly outperforms the HAR-based forecasts with an impressive relative R^2_{OOS} of 11.51%. The value-weighted relative R^2_{OOS}'s reported in Panel B evidence very similar patterns. The LASSO-based forecasts generally dominate the OLS-based forecasts, except for the very last subperiod, where it performs on par with the most general OLS-based model.

In light of the larger relative gains seemingly observed for the LASSO-based predictions during more volatile earlier time periods, we also perform a more detailed analysis pertaining to the recent Covid period. Although the data for the Covid period never enter our training and validation samples, LASSO is purposely designed to avoid overfitting the "noise" thereby hopefully attaining more robust predictions during that period as well. To corroborate this conjecture, Figure 8 presents a scatter plot of the average realized and predicted correlations

for each of the stocks in our sample during the peak six-month Covid period from March 2020, when the World Health Organization declared COVID-19 a pandemic, through August 2020 a half-year later. The reported average correlations for each of the stocks are simply calculated as the mean of all the pairwise correlations that contained a particular stock. As the figure shows, most of the LASSO-based correlation predictions are indeed closer to the 45-degree line than the HAR-based predictions, reaffirming that consistent with our previous out-of-sample evaluations LASSO typically produces the most accurate forecasts over this challenging time period as well.

6.2. Alternative features and machine learning techniques

The 25 features underlying our main results all involve various realized correlation-type measures. As discussed above, our use of these measures as our main features is naturally motivated by earlier findings in the financial econometrics literature. Meanwhile, a number of different economically motivated measures have also previously been proposed in the finance literature for capturing firm connections. Concretely, we consider the distance between two firms' headquarters as measured by their Zipcode distance (Parsons et al., 2020), economic activity as measured by text-based network industry classifications (Hoberg and Phillips, 2010, 2016), industry supply chain dependence (Menzly and Ozbas, 2010), common analyst coverage (Israelsen, 2016), common active mutual fund ownership (Antón and Polk, 2014), and common passive mutual fund ownership (Appel et al., 2016).[24] Table A.3 in the Appendix provides more detailed definitions and data sources.

To investigate whether any of these six additional economically-motivated firm-linkage variables may be used to further improve the out-of-sample correlation forecasts, we begin by succinctly summarizing the information in each by a series of simple dummies using the medians as cutoffs, the only exception being the dummy for text-based industry classification

[24] A number of other firm linkage measures have also been explored in the literature, such as stocks with similar value-growth labeling (Boyer, 2011) and firms with similar retail concentration (Kumar and Lee, 2006).

(*TNIC3*) which already takes a value of one if two firms are classified into the same industry and zero otherwise. Armed with these dummies, we then consider two new augmented feature sets, the first set consisting of our original 25 features plus the 6 dummies, and the second set consisting of the 25 main features plus the 150 additional features obtained by interacting each of the original features with the six new dummies. As such, this allows us to examine both first-order direct and second-order interactive effects.

In addition to these alternative feature sets, we also consider the use of alternative machine learning algorithms for the construction of the prediction models. In particular, while our main results suggest that LASSO is effective in reducing the number of parameters in the forecasting models through the use of an L_1-norm penalty, it assumes a linear relation between the forecasts and the features. It is possible that this is too restrictive and that other constraints and/or nonlinear relations may result in even better forecasts. Hence, to investigate this we implement the following four additional machine learning algorithms: Ridge Regression (Ridge), Elastic Net (ENet), Principal Component Regression (PCR), and Feed-forward Neural Networks (FNN). The first three of these allow us to explore different regions of the parameter space in linear models, while the last allows us to more thoroughly explore possible nonlinear relations and interactions.[25]

Table 6 reports the R^2_{OOS} relative to the simple HAR forecasts obtained by different combinations of the feature sets and machine learning algorithms. We purposely maintain the same feature standardization and training scheme underlying the results reported in Table 2 to allow for a direct comparison with the earlier results.[26] In line with the earlier results, the equal and value-weighted R^2_{OOS}'s reported in Panels A and B, respectively, are fairly similar, suggesting that our findings are not dominated by many small firms and/or a few large firms.

[25]For the Principal Component Regressions, to help increase the computational speed and prevent overfitting, we restrict the number of components used in the regressions to be no more than 20. For the Neural Network, we report the results based on a two-hidden-layer feed-forward network with 4 and 2 neurons. However, in unreported results, we also considered a three-hidden-layer network with 8, 4, and 2 neurons, and a four-hidden-layer network with 16, 8, 4, and 2 neurons.

[26]The inclusion of manually selected interacted terms in the largest feature set further facilitates a fair comparison among the linear models (i.e., LASSO, Ridge, ENet, and PCR) and the non-linear models (i.e., FNN), and helps pin down the contribution of the interactive effects.

Looking across the rows, for the LASSO, Ridge, and ENet-based predictions, we observe only minor increases in the R^2_{OOS}'s by adding the new firm-linkage dummies, whether directly or interactively. As a case in point, for the LASSO-based predictions, the equal-weighted relative R^2_{OOS} increases from 10.16% for the model based on our original 25 features, to just 10.24% and 10.35% for the models based on the new augment feature sets, indirectly suggesting that the relevant information about firm linkages for correlation prediction are already embedded in the time-series based realized features. Further corroborating this conjecture, for both the PCR and FNN-based predictions, the inclusion of the additional firm-linkage measures actually results in lower R^2_{OOS}'s compared to the prediction models based on our original 25 features only.

Looking across the columns, as expected the performance of ENet is roughly comparable to that of LASSO, while the performance of Ridge appears slightly inferior. In contrast to LASSO, which is based on the sparsity-encouraging L_1-penalty, Ridge relies on an L_2-penalty, and thus never enjoys the benefits of true dimension reduction. By comparison, PCR does benefit from dimension reduction, as long as the specific feature set can be well-represented by the selected principal components. However, when this is not the case, PCR tends to underperform LASSO, as evidenced by the R^2_{OOS}'s for the third and largest feature set. Interestingly, the more complicated-to-implement FNN-based models also typically underperform LASSO.

In sum, the results in Table 6 further underscore the superior performance of our simple-to-construct LASSO-based correlation forecasting models, both in terms of the use of alternative firm-linkage measures and more complex machine learning algorithms.

7. Conclusion

Exploiting techniques and ideas from the recent machine learning literature, we design new features and models explicitly geared toward correlation forecasting. Implementing the new models with a large panel of S&P 500 stocks, we document statistically significant

improvements in out-of-sample forecast accuracy, along with nontrivial economic gains from applying the new models in a wide range of practical applications. We further attribute the success of our preferred LASSO-based models to their ability to adapt to changing market conditions by allowing the many different features to dynamically enter and exit the predictions. It is, of course, possible that even better forecasting models could be constructed by the use of more complicated fitting algorithms, such as, e.g., tree models, together with additional economically motivated enlarged feature sets. By comparison, our current focus on linear models and algorithms is explicitly designed to allow for greater transparency and easier interpretability of the results. The same ideas and techniques developed here could also be used in the construction of forecasting models for other commonly used risk measures, including measures of precision and factor risk exposures.

Another promising avenue for future research would entail the use of realized correlation forecasts to study the correlation risk premium. Driessen, Maenhout, and Vilkov (2009) compares the variance risk premium (VRP) of a broadly defined market index and its constituents, and concludes that the significant VRP for the market is primarily driven by correlation risk since the VRPs for most individual stocks are not significantly different from zero. Other studies rely on correlation innovation (Krishnan, Petkova, and Ritchken, 2009), or the difference between option-implied and realized correlations (Driessen, Maenhout, and Vilkov, 2013; Mueller, Stathopoulos, and Vedolin, 2017; Bondarenko and Bernard, 2021) to further elucidate the origins of the correlation risk premium. Our new framework for more accurate realized correlation forecasting holds the promise of additional insights by explicating the economic drivers behind changes in correlation risks. Other potentially interesting applications include the use of our correlation forecasts to predict factor betas (Buss and Vilkov, 2012), and better evaluate hedge funds' ability to maintain market neutrality and assess their timing ability (Buraschi, Kosowski, and Trojani, 2014), to name but a few. We leave further work along all of these lines for future research.

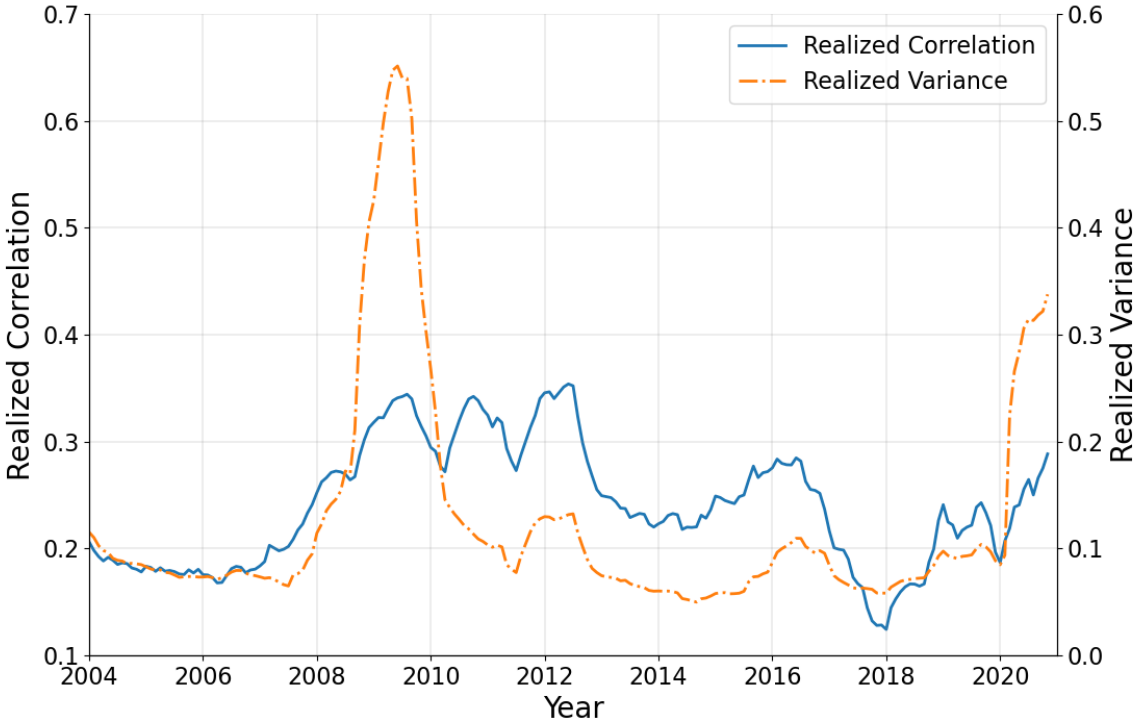


Fig. 1 Monthly realized correlations and variances

The figure plots the 12-month moving average of the cross-sectional means of the monthly realized correlations and realized variances for the S&P 500 stocks in our sample.

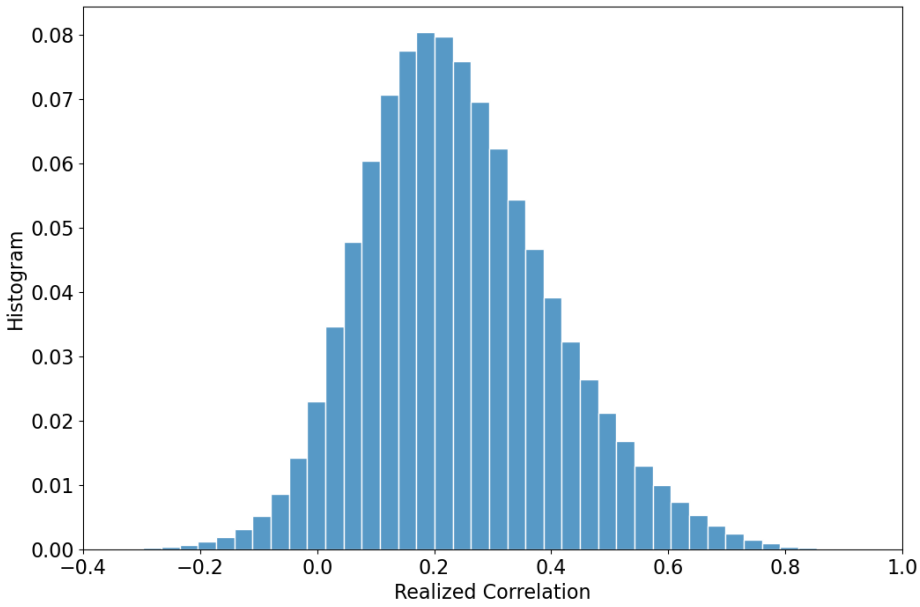


Fig. 2 Monthly realized correlations

The figure displays the histogram of monthly realized correlations for the S&P 500 stocks in our sample.

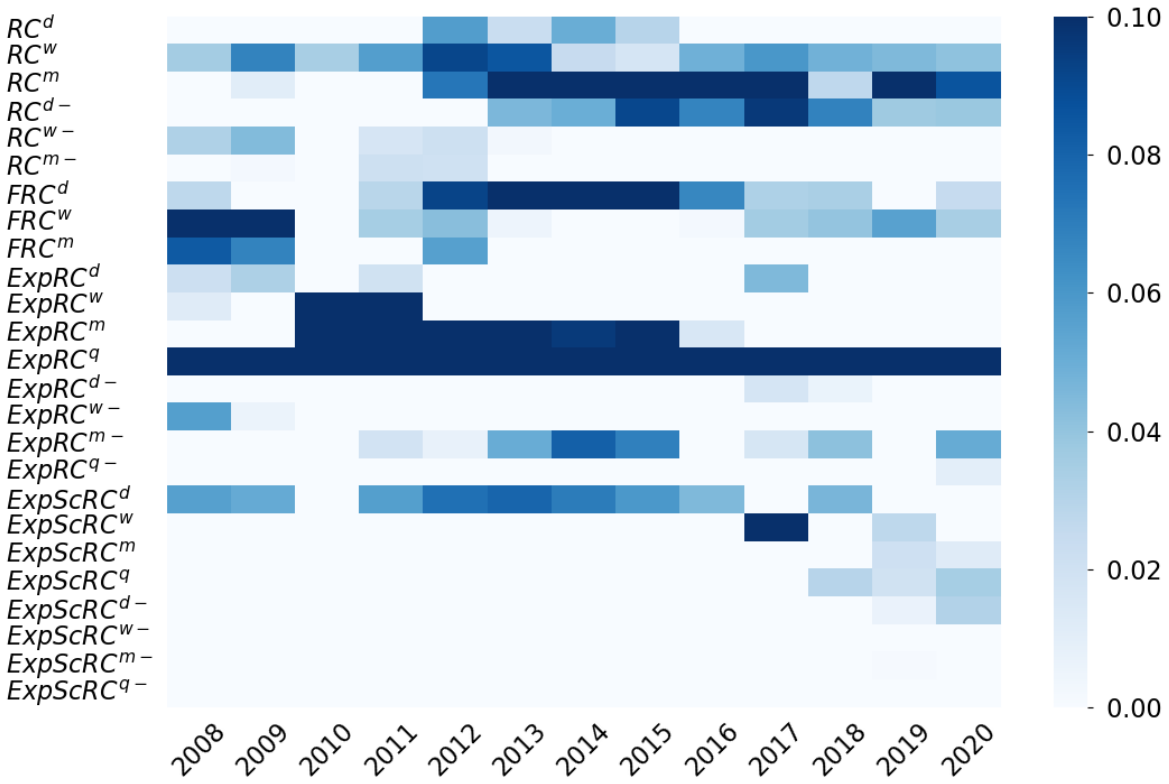
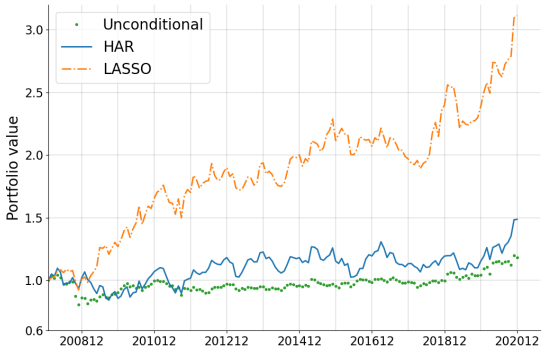


Fig. 3 Contribution of selected features

The figure displays the scaled absolute coefficients on the 25 main realized features used in the LASSO model. The columns denote each year in the testing sample, and the rows display the 25 features. The color gradients within each column indicate the most influential (dark blue) to the least influential (light blue) features, with white indicating features that are not selected.

Panel A: Equal-weighted



Panel B: Value-weighted

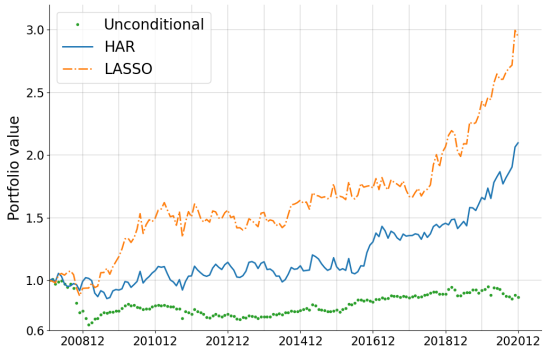


Fig. 4 Cumulative values for pairs trading strategies

The figure displays the cumulative gains of the pairs trading strategies based on the $RetDiff$ signals for different stock samples from February 2008 to December 2020. The samples for the pairs trading strategy include the full S&P 500 universe (Unconditional), the conditional sample based on the correlation predictions from the HAR model (HAR), and the conditional sample based on the correlation predictions from the LASSO-based model (LASSO). The $RetDiff$ signal is the return divergence between stocks and their pair portfolios, as defined in equation (10) in the main text.

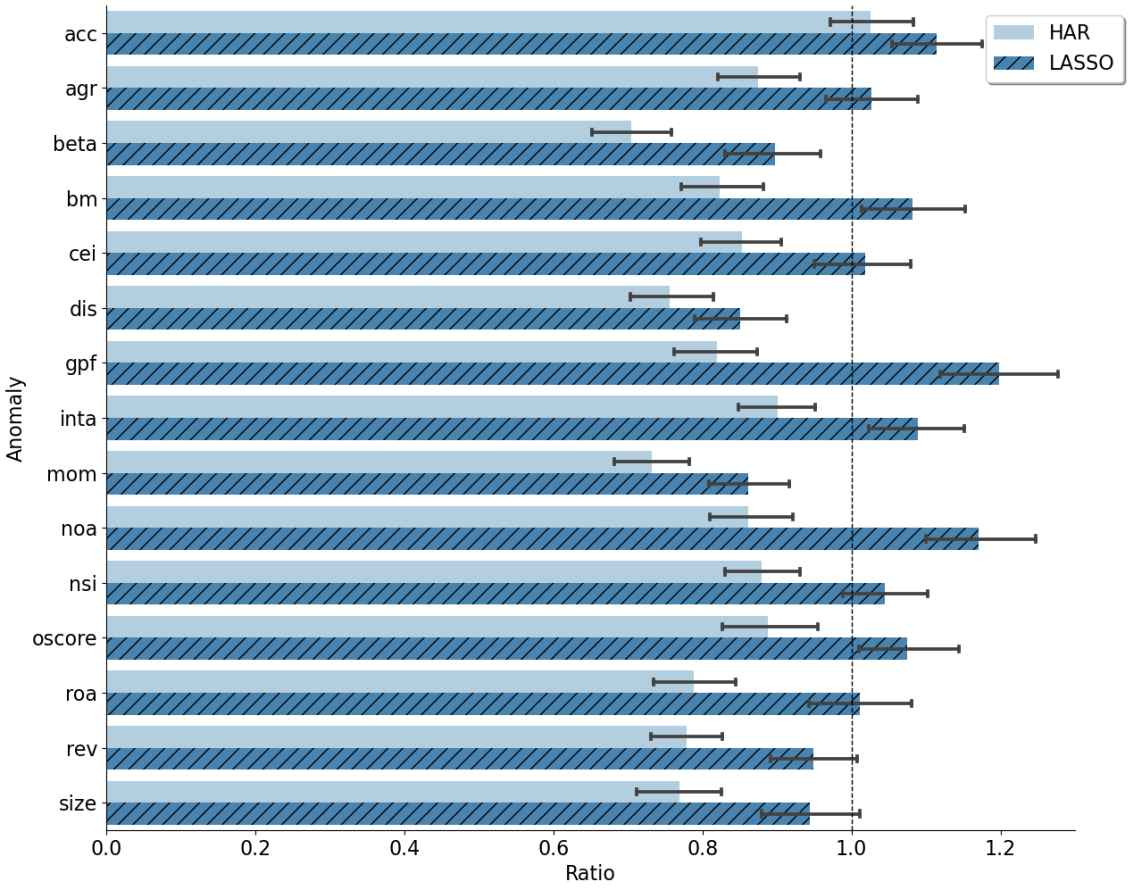


Fig. 5 Risk-targeting ratios of long-short strategies

The figure displays the ratios of the forecasted portfolio risk over the realized portfolio risk from the HAR and LASSO-based models for the 15 main long-short trading strategies discussed in the main text. The ratios are averaged over all testing samples according to equation (14). The black lines correspond to 95% bootstrapped confidence intervals.

	Mean Ret	St.Dev.	Sharpe Ratio	γ=2	γ=5	γ=10
HAR	10.27%	36.42%	0.36			
LASSO	10.90%	34.49%	0.48	0.77%	0.98%	1.37%

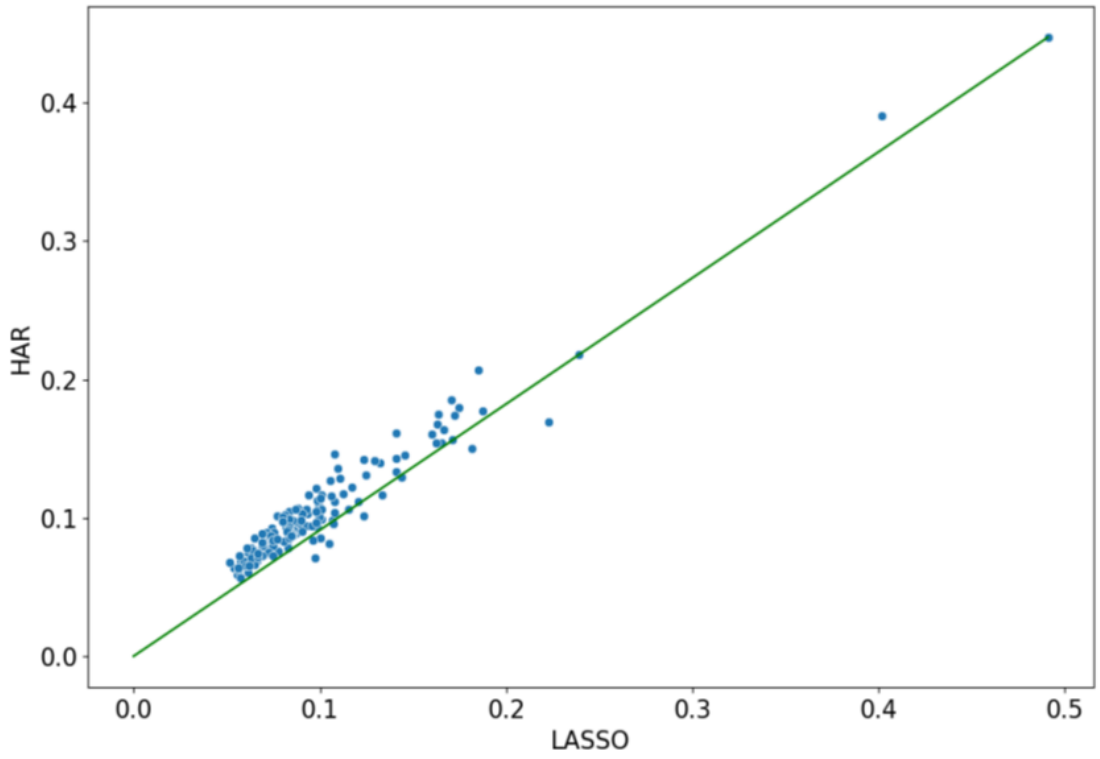


Fig. 6 GMV portfolio comparisons

The figure compares the performance of GMV portfolios for the HAR and LASSO-based forecasting models. The top panel reports the full sample average returns, standard deviations, and Sharpe ratios for the GMV portfolios, together with the utility gains (in annualized returns) of switching from HAR- to LASSO-based forecasts. The scatter plot displays the (annualized) monthly standard deviations from the two different models.

	Mean Ret	St.Dev.	Sharpe Ratio	Realized Beta	ϒ=2	ϒ=5	ϒ=10
HAR	12.13%	51.17%	0.26	-0.20 (-7.19)			
LASSO	13.14%	43.12%	0.39	0.05 (1.58)	1.81%	3.36%	6.07%

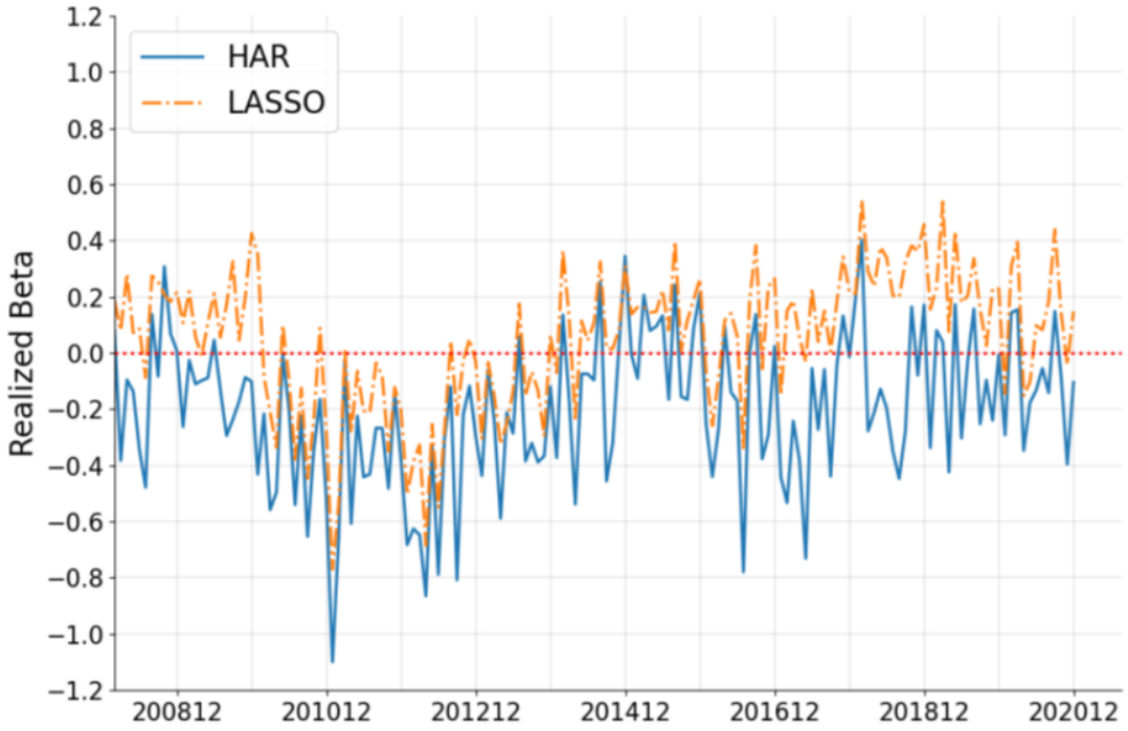


Fig. 7 Beta-neutral GMV portfolio comparisons

The figure compares the performance of beta-neutral GMV portfolios for the HAR and LASSO-based forecasting models. The top panel reports the full sample average returns, standard deviations, Sharpe ratios, and realized betas for the portfolios, together with the utility gains (in annualized returns) of switching from the HAR to the LASSO-based forecasts. The time-series plot displays the monthly realized betas for the beta-neutral GMV portfolios from the two different forecasting models.

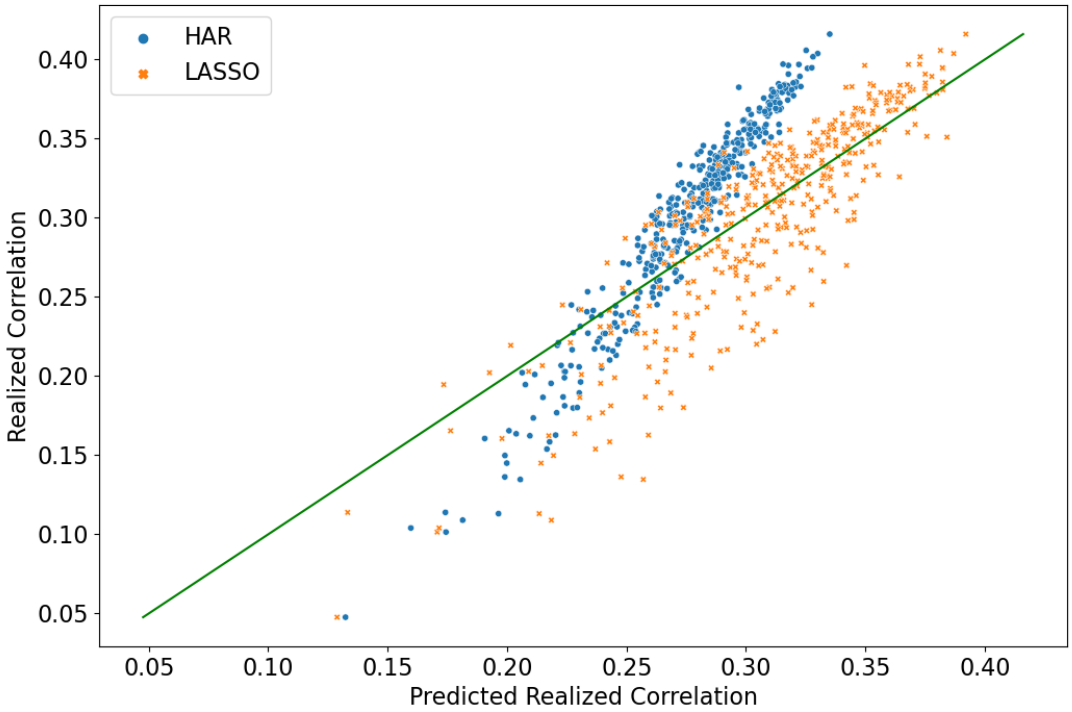


Fig. 8 Out-of-sample predictions during the peak of Covid

The figure shows for each stock in our sample the average realized correlations (y-axis) against the average predicted correlations (x-axis) from the HAR model and the LASSO-based model between March 2020 and August 2020. The average correlation for a stock is calculated as the mean of all the pairwise correlations that contain the stock.

Table 1 Descriptive statistics for realized correlation features

The table reports the descriptive statistics of all realized correlation features. The sample consists of 417 stocks listed on NYSE/AMEX/NASDAQ that have ever been included in the S&P 500 index and have full historical quotations over the period from January 2000 to December 2020 with share code 10 or 11, price between \$1 and \$1000, and daily number of trades greater than or equal to 100. Superscripts d, w, m, and q are abbreviations of daily, weekly, monthly, and quarterly construction intervals or center-of-mass. RC^h and RC^{h-} $(h = d, w, m)$ denote the daily, weekly, and monthly realized correlation and negative semicorrelation, respectively. FRC^h $(h = d, w, m)$ is the daily, weekly, and monthly factor-driven realized correlation. $ExpRC^h$ and $ExpRC^{h-}$ $(h = d, w, m, q)$ are the exponential realized correlation and negative semicorrelation calculated from exponentially weighted moving average of past 500-day realized covariances and negative semicovariances using the corresponding center-of-mass h. $ExpScRC^h$ and $ExpScRC^{h-}$ $(h = d, w, m, q)$ are the sector-specific exponential realized correlations and semicorrelations with center-of-mass h.

Variable	Mean	Std	Skewness	Kurtosis	Min	P1	P25	Median	P75	P99	Max	AR(1)	AR(5)	AR(21)	AR(63)
RC^d	0.26	0.30	-0.33	-0.21	-0.98	-0.49	0.06	0.28	0.48	0.85	1.00	0.09	0.03	0.03	-0.04
RC^w	0.25	0.21	0.06	0.02	-0.93	-0.22	0.11	0.25	0.39	0.74	1.00	0.20	0.09	0.10	-0.05
RC^m	0.24	0.16	0.36	0.25	-0.89	-0.10	0.13	0.23	0.34	0.66	0.98	0.40	0.18	0.11	-0.04
RC^{d-}	0.46	0.22	0.07	-0.83	0.00	0.04	0.29	0.46	0.63	0.91	1.00	0.08	0.03	0.02	-0.03
RC^{w-}	0.44	0.16	0.21	-0.32	0.00	0.11	0.32	0.43	0.55	0.81	1.00	0.16	0.07	0.08	-0.03
RC^{m-}	0.41	0.13	0.31	0.07	0.01	0.14	0.33	0.41	0.50	0.74	0.98	0.28	0.14	0.11	-0.03
FRC^d	0.24	0.22	0.91	1.76	-1.00	-0.14	0.09	0.20	0.35	0.98	1.00	0.32	0.17	0.09	-0.08
FRC^w	0.22	0.17	0.97	1.51	-1.00	-0.07	0.10	0.19	0.31	0.75	1.00	0.46	0.26	0.14	-0.10
FRC^m	0.21	0.15	0.94	1.53	-1.00	-0.05	0.10	0.18	0.29	0.66	1.00	0.60	0.33	0.14	-0.09
$ExpRC^d$	0.25	0.22	-0.03	0.04	-0.95	-0.26	0.10	0.25	0.40	0.75	0.99	0.17	0.07	0.08	-0.05
$ExpRC^w$	0.24	0.17	0.34	0.31	-0.88	-0.11	0.12	0.23	0.34	0.67	0.97	0.40	0.15	0.13	-0.05
$ExpRC^m$	0.24	0.14	0.52	0.43	-0.69	-0.03	0.14	0.22	0.32	0.61	0.93	0.79	0.35	0.12	-0.15
$ExpRC^q$	0.24	0.12	0.54	0.37	-0.46	0.00	0.16	0.23	0.32	0.58	0.93	0.93	0.64	0.17	-0.30
$ExpRC^{d-}$	0.44	0.17	0.21	-0.41	0.00	0.11	0.32	0.43	0.55	0.83	0.99	0.14	0.07	0.07	-0.03
$ExpRC^{w-}$	0.41	0.13	0.32	0.08	0.01	0.13	0.32	0.41	0.50	0.75	0.98	0.33	0.13	0.12	-0.04
$ExpRC^{m-}$	0.40	0.11	0.38	0.25	0.03	0.17	0.32	0.39	0.47	0.70	0.96	0.75	0.30	0.11	-0.12
$ExpRC^{q-}$	0.40	0.10	0.42	0.22	0.04	0.19	0.33	0.39	0.46	0.66	0.95	0.91	0.61	0.16	-0.25
$ExpScRC^d$	0.04	0.12	2.98	8.30	0.00	0.00	0.00	0.00	0.00	0.54	0.82	0.40	0.18	0.21	-0.10
$ExpScRC^w$	0.04	0.11	2.98	8.31	0.00	0.00	0.00	0.00	0.00	0.51	0.80	0.56	0.26	0.22	-0.09
$ExpScRC^m$	0.04	0.11	2.90	7.71	0.00	0.00	0.00	0.00	0.00	0.49	0.75	0.86	0.46	0.17	-0.22
$ExpScRC^q$	0.04	0.11	2.83	7.10	0.00	0.00	0.00	0.00	0.00	0.49	0.69	0.96	0.73	0.24	-0.38
$ExpScRC^{d-}$	0.06	0.17	2.48	4.55	0.00	0.00	0.00	0.00	0.00	0.64	0.90	0.40	0.20	0.19	-0.10
$ExpScRC^{w-}$	0.06	0.16	2.48	4.60	0.00	0.00	0.00	0.00	0.00	0.61	0.85	0.53	0.25	0.25	-0.09
$ExpScRC^{m-}$	0.06	0.15	2.48	4.58	0.00	0.00	0.00	0.00	0.00	0.59	0.79	0.85	0.46	0.20	-0.22
$ExpScRC^{q-}$	0.06	0.15	2.47	4.51	0.00	0.00	0.00	0.00	0.00	0.57	0.75	0.96	0.74	0.28	-0.38

Table 2 Out-of-sample predictions

The table reports the out-of-sample performance for OLS-based correlation forecasting models and LASSO-based correlation forecasting model. The sample consists of 417 stocks listed on NYSE/AMEX/NASDAQ that have ever been included in the S&P 500 index and have full historical quotations over the period from January 2000 to December 2020 with share code 10 or 11, price between \$1 and \$1,000, and daily number of trades greater than or equal to 100. Superscripts d, w, m, and q are abbreviations of daily, weekly, monthly, and quarterly construction intervals. RC^h and RC^{h-} $(h = d, w, m)$ denote the daily, weekly, and monthly realized correlation and negative realized correlation, respectively. FRC^h $(h = d, w, m)$ is the daily, weekly, and monthly factor-driven realized correlation. $ExpRC^h$ and $ExpRC^{h-}$ $(h = d, w, m, q)$ are the exponential realized correlation and negative semicorrelation calculated from exponentially weighted moving average of past 500-day realized covariances and negative semicovariances using the corresponding center-of-mass h. $ExpScRC^h$ and $ExpScRC^{h-}$ $(h = d, w, m, q)$ are the sector-specific exponential realized correlations and semicorrelations with center-of-mass h. Our correlation forecasting models include three OLS-based models: SHAR, SHAR-F, and SHAR-F-Exp, and the LASSO model. Panel A reports $R^{2,EW}_{OOS}$ and $R^{2,VW}_{OOS}$ relative to HAR for each model using the entire panel of stock pairs according to equations (7) and (8). Panel B and C report the modified DM t-statistics and value-weighted DM t-statistics for pairwise comparisons among models, respectively.

Panel A: R^2_{OOS} relative to HAR

	Model	Feature Set	Equal-weighted	Value-weighted
(1)	SHAR	3 RC^h + 3 RC^{h-} (# of features = 6)	0.22%	0.11%
(2)	SHAR-F	3 RC^h + 3 RC^{h-} + 3 FRC^h (# of features = 9)	1.71%	1.30%
(3)	SHAR-F-Exp	3 RC^h + 3 RC^{h-} + 3 FRC^h + 4 $ExpRC^h$ + 4 $ExpRC^{h-}$ + 4 $ExpScRC^h$ + 4 $ExpScRC^{h-}$ (# of features = 25)	9.82%	7.31%
(4)	LASSO	All 25 main features	10.16%	8.05%

Panel B: DM t-statistics (equal-weighted)

	Model	HAR	(1)	(2)	(3)
(1)	SHAR	11.55			
(2)	SHAR-F	29.32	27.58		
(3)	SHAR-F-Exp	39.08	39.84	35.24	
(4)	LASSO	47.70	48.93	43.43	6.31

Panel C: DM t-statistics (value-weighted)

	Model	HAR	(1)	(2)	(3)
(1)	SHAR	4.99			
(2)	SHAR-F	13.56	13.41		
(3)	SHAR-F-Exp	16.21	16.29	15.51	
(4)	LASSO	17.85	17.91	17.41	8.99

Table 3 Performance of pairs trading strategies

The table reports the performance of the pairs trading strategy based on $RetDiff$ signals using different stock samples. The full evaluation period is from February 2008 to December 2020. The samples for pairs trading strategy are S&P 500 universe (unconditional sample) and reduced samples based on HAR and LASSO correlation predictions. $RetDiff$ signal is the return divergence between stocks and their pair portfolios defined in equation (10). Panels A and B report the annualized equal-weighted and value-weighted monthly returns of the quintile portfolios sorted by $RetDiff$ with different samples. By the end of each month t, we sort stocks on $RetDiff$ into five quintile portfolios. We then compute the difference between predicted and historical correlations of a stock's pair portfolio, $\Delta Corr^{\theta}$, based on HAR and LASSO correlation predictions. Lastly we calculate realized monthly returns in month $t+1$ for $RetDiff$-sorted portfolios with the unconditional sample and restricted samples of stocks in quintile 5 sorted by $\Delta Corr^{\theta}$. The column labeled "HML" reports the difference in returns between Portfolio 5 and Portfolio 1, with t-statistics in parentheses. Panel C reports the estimated regression coefficients and t-statistics (in-parentheses) from Fama-MacBeth cross-sectional regressions predicting one-month ahead stock returns using $RetDiff$ defined as discrete values of [1, 2, 3, 4, 5] according to sorted quintiles each month. The control variables include *Beta*, *Size*, *BM*, and *Mom*.

Panel A: Equal-weighted portfolio sorted by return divergence

	1 (Low)	2	3	4	5 (High)	HML
Unconditional	6.92%	5.75%	7.09%	6.45%	8.07%	1.15% (0.47)
HAR	9.53%	6.44%	9.25%	7.90%	13.16%	3.63% (0.88)
LASSO	3.50%	5.12%	6.08%	5.67%	12.84%	9.34% (2.30)

Panel B: Value-weighted portfolio sorted by return divergence

	1 (Low)	2	3	4	5 (High)	HML
Unconditional	6.05%	4.58%	6.05%	4.76%	4.86%	-1.20% (-0.45)
HAR	6.42%	6.63%	9.02%	7.90%	12.56%	6.14% (1.60)
LASSO	1.90%	5.68%	6.63%	6.28%	10.75%	8.85% (2.20)

Panel C: Fama-MacBeth regressions

	Unconditional		HAR		LASSO	
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	0.50	4.42	0.55	5.79	0.13	6.52
	(1.28)	(3.78)	(1.18)	(2.91)	(0.31)	(3.60)
RetDiff	0.03	0.04	0.08	0.12	0.14	0.16
	(0.54)	(0.97)	(1.02)	(1.80)	(1.87)	(2.33)
Controls	No	Yes	No	Yes	No	Yes
Adj-R^2	0.59%	12.01%	0.92%	11.73%	1.06%	12.82%
N	64,635	64,635	13,020	13,020	13,020	13,020

Table 4 Equity premium predictions

The table reports the OLS predictive regression results using different average correlation measures ($AvgCorr^{\theta}$) to forecast future market excess returns. $AvgCorr^{\theta}$ is calculated according to equation (12) based on a naive model using the lagged correlation as predictors (denoted RC in the table), the HAR model, and the LASSO-based model. The dependent variable is the monthly excess return of the CRSP value-weighted index. The sample spans from February 2008 to December 2020. The control variables include the dividend-price ratio (*dp*), earnings-price ratio (*ep*), book-to-market ratio (*bm*), net equity expansion (*ntis*), treasury-bill rate (*tbl*), term spread (*tms*), default spread (*dfy*), and stock market variance (*svar*), and *t*-statistics are reported in parentheses.

	Panel A: $AvgCorr^{EW}$						Panel B: $AvgCorr^{VW}$					
	(1)	(2)	(3)	(4)	(5)	(6)	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	0.00	-0.02	-0.02	0.54	0.49	0.53	0.00	-0.02	-0.03	0.55	0.50	0.56
	(0.05)	(-0.94)	(-1.41)	(1.53)	(1.33)	(1.52)	(0.18)	(-0.86)	(-1.49)	(1.54)	(1.37)	(1.63)
AvgCorr												
RC	0.03			0.03			0.03			0.02		
	(0.88)			(0.52)			(0.75)			(0.37)		
HAR		0.11			0.08			0.10			0.06	
		(1.47)			(0.73)			(1.39)			(0.61)	
LASSO			0.13			0.24			0.13			0.25
			(2.00)			(2.40)			(2.08)			(2.66)
dp				0.12	0.11	0.13				0.12	0.11	0.14
				(1.65)	(1.49)	(1.77)				(1.66)	(1.52)	(1.90)
ep				-0.00	-0.00	-0.01				-0.00	-0.00	-0.00
				(-0.17)	(-0.27)	(-0.31)				(-0.14)	(-0.23)	(-0.21)
bm				-0.12	-0.13	-0.15				-0.12	-0.13	-0.13
				(-0.83)	(-0.88)	(-0.99)				(-0.82)	(-0.85)	(-0.93)
ntis				0.22	0.20	0.11				0.23	0.22	0.10
				(0.65)	(0.62)	(0.33)				(0.69)	(0.66)	(0.31)
tbl				-0.99	-0.91	-0.63				-0.99	-0.93	-0.67
				(-1.36)	(-1.25)	(-0.87)				(-1.36)	(-1.27)	(-0.92)
tms				0.11	0.10	0.09				0.12	0.11	0.09
				(0.87)	(0.83)	(0.73)				(0.92)	(0.87)	(0.73)
dfy				-2.99	-2.91	-4.84				-2.89	-2.83	-5.03
				(-1.74)	(-1.74)	(-2.62)				(-1.70)	(-1.70)	(-2.74)
svar				-0.18	-0.15	-0.23				-0.17	-0.17	-0.41
				(-0.31)	(-0.27)	(-0.41)				(-0.29)	(-0.30)	(-0.73)
Adj-R^2	-0.15%	0.74%	1.91%	1.76%	1.94%	5.33%	-0.29%	0.60%	2.12%	1.67%	1.82%	6.16%
N	155	155	155	155	155	155	155	155	155	155	155	155

Table 5 Out-of-sample predictions for different subsamples

The table reports the out-of-sample performance for OLS and LASSO-based correlation forecasting models for the three subsample periods: 2008-2011, 2012-2015, and 2016-2020. The sample consists of 417 stocks listed on NYSE/AMEX/NASDAQ that have ever been included in the S&P 500 index and have full historical quotations over the period from January 2000 to December 2020 with share code 10 or 11, price between \$1 and \$1,000, and daily number of trades greater than or equal to 100. Superscripts d, w, m, and q are abbreviations of daily, weekly, monthly, and quarterly construction intervals. RC^h and RC^{h-} $(h = d, w, m)$ denote the daily, weekly, and monthly realized correlation and negative realized correlation, respectively. FRC^h $(h = d, w, m)$ is the daily, weekly, and monthly factor-driven realized correlation. $ExpRC^h$ and $ExpRC^{h-}$ $(h = d, w, m, q)$ are the exponential realized correlation and negative semicorrelation calculated from exponentially weighted moving average of past 500-day realized covariances and negative semicovariances using the corresponding center-of-mass h. $ExpScRC^h$ and $ExpScRC^{h-}$ $(h = d, w, m, q)$ are the sector-specific exponential realized correlations and semicorrelations with center-of-mass h. Our correlation forecasting models include three OLS-based models: SHAR, SHAR-F, and SHAR-F-Exp, and the LASSO model. Panels A and B report $R^{2,EW}_{OOS}$ and $R^{2,VW}_{OOS}$ relative to the HAR model, respectively.

	Model	Feature set	R^2_{OOS} relative to HAR		
		Panel A: Equal-weighted			
			2008-2011	2012-2015	2016-2020
(1)	SHAR	3 RC^h + 3 RC^{h-} (# of features = 6)	0.12%	0.33%	0.23%
(2)	SHAR-F	3 RC^h + 3 RC^{h-} + 3 FRC^h (# of features = 9)	2.34%	0.64%	1.97%
(3)	SHAR-F-Exp	3 RC^h + 3 RC^{h-} + 3 FRC^h + 4 $ExpRC^h$ + 4 $ExpRC^{h-}$ + 4 $ExpScRC^h$ + 4 $ExpScRC^{h-}$ (# of features = 25)	6.95%	9.95%	11.89%
(4)	LASSO	All 25 main features	7.87%	10.70%	11.51%
		Panel B: Value-weighted			
			2008-2011	2012-2015	2016-2020
(1)	SHAR	3 RC^h + 3 RC^{h-} (# of features = 6)	0.08%	0.25%	0.05%
(2)	SHAR-F	3 RC^h + 3 RC^{h-} + 3 FRC^h (# of features = 9)	2.24%	0.04%	1.44%
(3)	SHAR-F-Exp	3 RC^h + 3 RC^{h-} + 3 FRC^h + 4 $ExpRC^h$ + 4 $ExpRC^{h-}$ + 4 $ExpScRC^h$ + 4 $ExpScRC^{h-}$ (# of features = 25)	3.66%	9.40%	8.47%
(4)	LASSO	All 25 main features	5.76%	10.10%	8.31%

Table 6 Out-of-sample predictions for alternative fitting procedures

The table reports the out-of-sample performance of different machine learning correlation forecasting models using alternative feature sets. The sample consists of 417 stocks listed on NYSE/AMEX/NASDAQ that have ever been included in the S&P 500 index and have full historical quotations over the period from January 2000 to December 2020 with share code 10 or 11, price between \$1 and \$1,000, and the daily number of trades greater than or equal to 100. The features include 25 predictors used in the main analysis, six dummies based on firm-linkage variables, and 150 feature-dummy interactive terms. Our correlation forecasting models include LASSO, Ridge Regression (Ridge), Elastic Net (ENet), Principal Component Regression (PCR), and two-hidden-layer Feed-Forward Neural Network (FNN). Panels A and B report $R^{2,EW}_{OOS}$ and $R^{2,VW}_{OOS}$ relative to HAR, as defined in equations (7) and (8), respectively, for each model using different sets of features.

Feature set	R^2_{OOS} relative to HAR				
Panel A: Equal-weighted					
	LASSO	Ridge	ENet	PCR	FNN
All 25 main features	10.16%	9.83%	10.14%	10.44%	10.12%
All 25 main features + 6 dummies (# of features = 31)	10.24%	9.96%	10.19%	9.61%	9.97%
All 25 main features + 150 feature × dummy combinations (# of features = 175)	10.35%	9.95%	10.31%	8.76%	9.88%
Panel B: Value-weighted					
	LASSO	Ridge	ENet	PCR	FNN
All 25 main features	8.05%	7.31%	8.07%	8.31%	7.56%
All 25 main features + 6 dummies (# of features = 31)	8.05%	7.38%	8.09%	7.66%	6.98%
All 25 main features + 150 feature × dummy combinations (# of features = 175)	8.20%	7.54%	8.24%	7.68%	7.02%

Appendix

Correlation signature plot

Following Andersen, Bollerslev, Diebold, and Labys (2000), Figure A.1 plots the sample mean correlations averaged across time and stocks as a function of different sampling frequencies ranging from 1-minute to 1-hour. As the figure shows, the averaged realized correlations increase from the 1- to 10-minute sampling frequency, but appear to flatten out at around the 15-minute frequency, underscoring the soundness of said sampling frequency used in the construction of our monthly realized correlation measures.

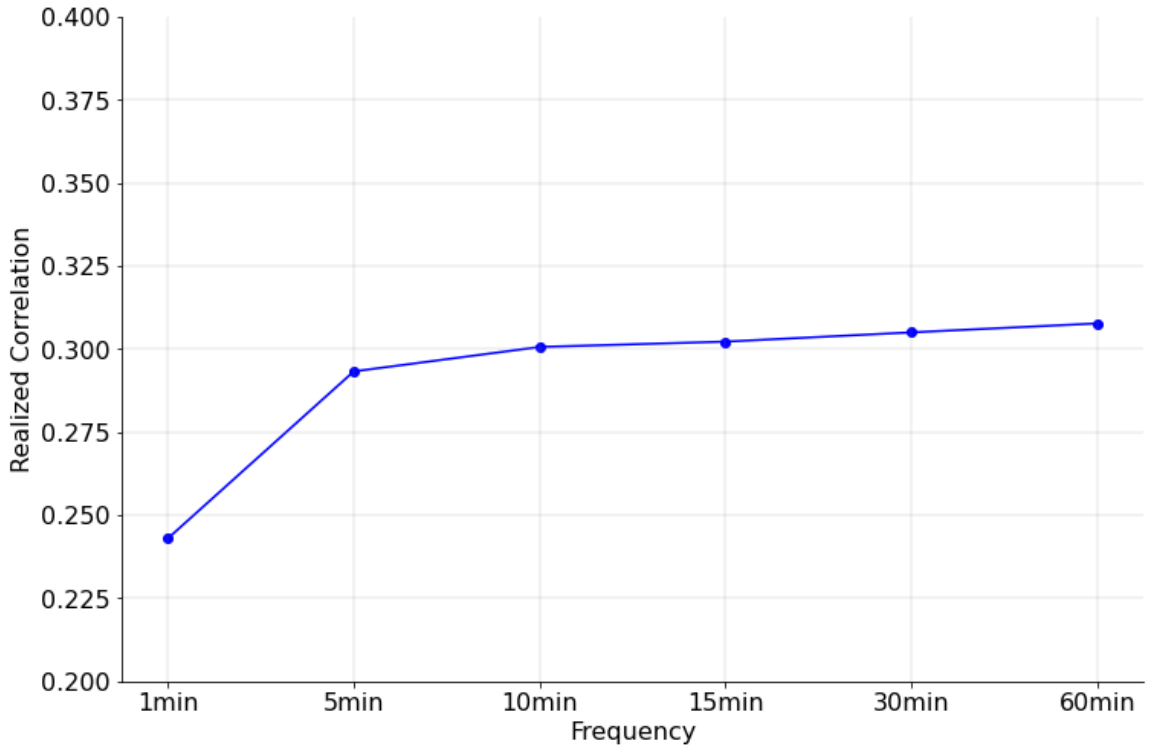


Fig. A.1 Signature plots for monthly realized correlation

This figure shows the mean value of monthly realized correlation averaged across stocks and time for different sampling frequencies.

Anomaly characteristics

Table A.1 Descriptive statistics for anomaly characteristics

This table reports the mean, standard deviation, and quantiles of 15 representative anomaly characteristics including 11 mispricing anomalies of Stambaugh et al. (2012) and *Beta*, *Size*, *Book-to-market ratio*, and *Reversal*. The sample consists of 417 stocks listed on NYSE/AMEX/NASDAQ that have ever been included in the S&P 500 index and have full historical quotations over the period from January 2000 to December 2020 with share code 10 or 11, price between 1 and 1000, and daily number of trades greater than or equal to 100.

Variable	Acronym	Mean	Std	P1	P25	Median	P75	P99
Accruals	acc	0.00	0.04	-0.12	-0.01	0.00	0.02	0.11
Asset growth	agr	0.10	0.25	-0.30	-0.01	0.05	0.13	1.29
Beta	beta	1.04	0.51	0.12	0.67	0.97	1.31	2.63
Book-to-market	bm	0.47	0.42	-0.09	0.22	0.37	0.62	1.82
Composite equity issues	cei	-0.08	0.23	-0.75	-0.10	-0.06	-0.03	0.36
Distress	dis	-6.50	5.41	-8.57	-7.42	-6.86	-6.01	0.50
Gross profitability	gpf	0.30	0.23	-0.01	0.12	0.26	0.42	1.02
Investment-to-assets	inta	0.06	3.70	-0.17	0.01	0.03	0.06	0.39
Momentum	mom	0.13	0.37	-0.61	-0.06	0.11	0.28	1.31
Net operating assets	noa	0.53	0.35	-0.20	0.36	0.54	0.67	1.53
Net stock issues	nsi	0.13	0.93	-0.15	-0.03	0.00	0.01	3.09
O-score	oscore	-3.91	1.60	-7.64	-4.78	-3.95	-3.16	0.77
Return on assets	roa	0.01	0.02	-0.07	0.00	0.01	0.03	0.08
Reversal	rev	0.01	0.10	-0.25	-0.03	0.01	0.06	0.28
Size	size	16.20	1.24	13.23	15.36	16.21	17.04	19.09

Additional anomaly characteristics

Table A.2 Descriptive statistics for additional firm characteristics

This table reports the mean, standard deviation, and quantiles of 15 alternative firm characteristics. The firm characteristics are another set of representative anomalies including Abnormal earnings announcement return (*abr*), Abnormal earnings announcement volume (*aeavol*), Change in 6-month momentum (*chmom*), Change in shares outstanding (*chcsho*), Current ratio (*currat*), Earnings to price (*ep*), Employee growth rate (*hire*), Expected growth (*eg*), Industry momentum (*indmom*), Industry-adjusted change in profit margin (*chpmia*), Investment (*invest*), Liquidity (*liq*), Long-term reversals (*lrv*), Residual variance (*rvr*), and Sales growth (*sgr*). The sample includes all 417 stocks listed on NYSE/AMEX/NASDAQ that have ever been included in the S&P 500 index and have full historical quotations over the period from January 2000 to December 2020.

Variable	Acronym	Mean	Std	P1	P25	Median	P75	P99
Abnormal earnings announcement return	abr	0.00	0.02	-0.05	-0.01	0.00	0.01	0.05
Abnormal earnings announcement volume	aeavol	0.87	0.96	-0.35	0.26	0.65	1.20	4.50
Change in 6-month momentum	chmom	0.01	0.37	-0.86	-0.17	-0.01	0.17	1.08
Change in shares outstanding	chcsho	0.04	0.22	-0.14	-0.02	0.00	0.01	1.05
Current ratio	currat	2.57	4.65	0.50	1.09	1.53	2.34	24.58
Earnings to price	ep	0.03	0.22	-0.56	0.03	0.05	0.07	0.17
Employee growth rate	hire	0.04	0.17	-0.38	-0.02	0.02	0.08	0.72
Expected growth	eg	0.00	0.02	-0.05	0.00	0.00	0.01	0.05
Industry momentum	indmom	0.12	0.29	-0.48	-0.04	0.11	0.24	1.11
Industry-adjusted change in profit margin	chpmia	0.52	7.43	-15.81	-0.17	0.00	0.12	37.83
Investment	invest	1.00	0.45	0.30	0.85	0.98	1.13	1.99
Liquidity	liq	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Long-term reversals	lrv	0.33	0.72	-0.76	-0.03	0.24	0.54	2.71
Residual variance	rvr	0.04	0.02	0.02	0.02	0.03	0.04	0.11
Sales growth	sgr	0.08	0.22	-0.44	0.00	0.06	0.13	0.83

Risk-targeting ratios for additional anomaly portfolios

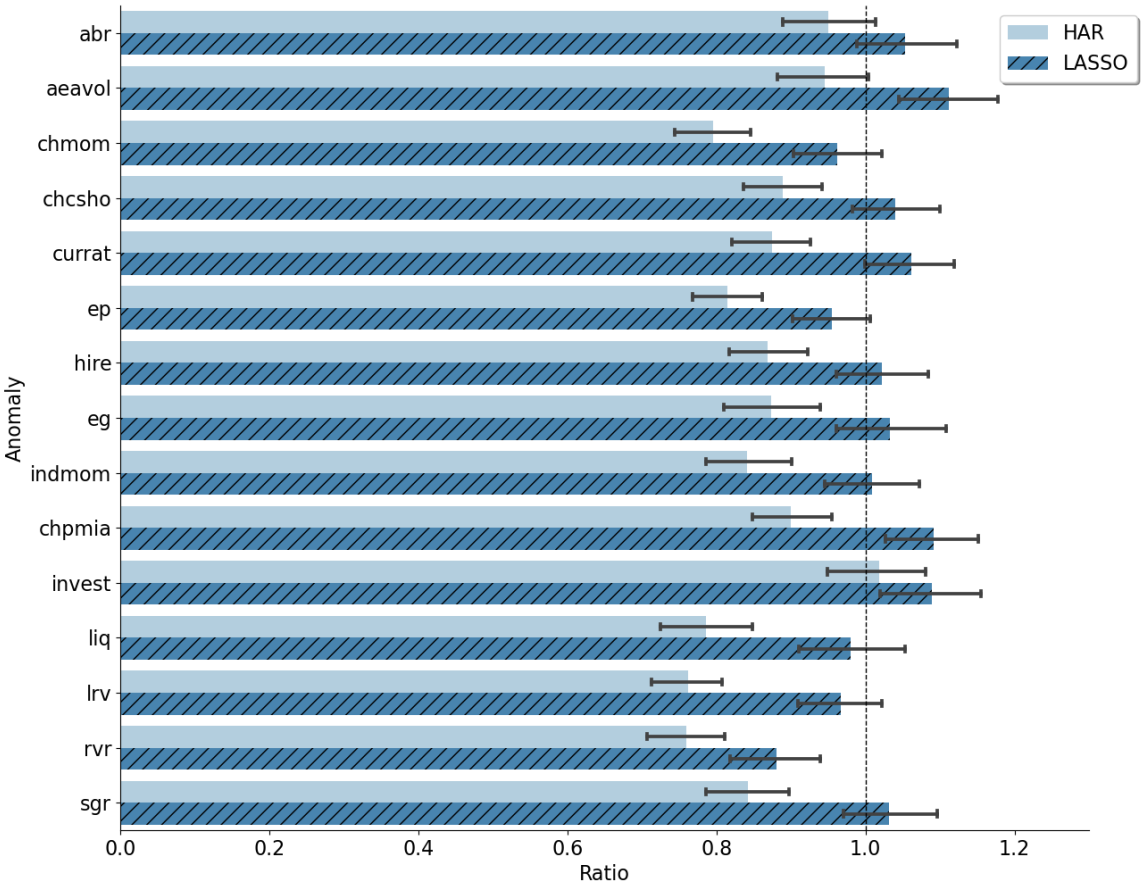


Fig. A.2 Risk-targeting ratios of additional long-short strategies

This figure displays the average ratios of forecasted portfolio risk over realized portfolio risk under HAR and LASSO models across 15 alternative long-short trading strategies defined in Table A.2. The average realized ratios are calculated over all testing samples according to equation (14). 95% bootstrapped confidence intervals are included.

Correcting non-positive definite matrices

The optimal solution of the GMV portfolio requires the input covariance matrices to be positive definite, or equivalently the correlation matrices formed by the pairwise correlation forecast must be positive definite. The simple HAR-based correlation matrix forecasts, of course, are always positive-definite since they are convex combinations of lagged daily, weekly, and monthly realized correlation matrices, each of which is positive-definite by construction. However, some of the inputs for the LASSO-based correlation matrix forecasts (e.g., the negative realized semicorrelation matrices) are not positive-definite, and the predicted correlation matrices are therefore not guaranteed to be positive definite. This occurs for about 10% of the forecasts in our sample. To address this issue, we apply a simple convexity correction on any non-positive-definite correlation matrix prediction:

$$\widehat{R}_t^{LASSO*} = a\widehat{R}_t^{HAR} + (1-a)\widehat{R}_t^{LASSO},$$

where the scalar a denotes the weight placed on the corresponding positive definite HAR-model forecasts. We choose the minimum value of $a > 0$ such that the resulting $\widehat{R}_t^{LASSO*}$ is positive definite, based on a threshold of 0.1 for the smallest eigenvalue to obtain a stable inverse.[27] Importantly, however, our GMV-related model comparison results reported in Figure 6 remain robust to the exclusion of the months in the sample for which the unadjusted LASSO-based forecasts are non-positive-definite, underscoring that the superior performance is not driven by this convexity correction.

[27]Relatedly, Shi et al. (2020) propose shrinking the sample eigenvalues of the inverse covariance matrix used in GMV construction. The recent study by Archakov and Hansen (2021) provides an alternative more complicated parameterization of the correlation matrix based on a multivariate extension of the traditional Fisher transform that automatically guarantees positive definiteness; see also Archakov et al. (2020).

Traditional firm-linkage measures

Table A.3 Firm linkages

The table provides the definitions of the 6 additional firm-linkage variables investigated in Section 6.2, along with their data sources.

Variable	Definition	Data source
ZipDist	Zip code distance between two firms' headquarters	NBER ZIP Code Distance Database Compustat
TNIC3	Text-based Network Industry Classifications based on firm pairwise similarity scores from text analysis of firm 10-K product descriptions	Hoberg and Phillips Data Library
IndSuppDep	Industry supply chain dependence measured by fraction of industry-by-industry purchases from input-output tables	Bureau of Economic Analysis
CmnAnalys	Common analyst coverage as # of common analysts following the stock pair over # of total unique analysts	I/B/E/S
CmnActOwn	Common active mutual fund ownership defined as the total dollar value of a stock pair held by common active mutual funds over the total dollar value of shares outstanding for the stock pair	CRSP Mutual Fund Database
CmnPssOwn	Common passive mutual fund ownership defined as the total dollar value of a stock pair held by common passive mutual funds over the total dollar value of shares outstanding for the stock pair	CRSP Mutual Fund Database

References

Andersen, T. G., Bollerslev, T., Diebold, F. X., Labys, P., 2000. Great realizations. Risk 13, 105–108.

Andersen, T. G., Bollerslev, T., Diebold, F. X., Labys, P., 2003. Modeling and forecasting realized volatility. Econometrica 71, 579–625.

Antón, M., Polk, C., 2014. Connected stocks. Journal of Finance 69, 1099–1127.

Appel, I. R., Gormley, T. A., Keim, D. B., 2016. Passive investors, not passive owners. Journal of Financial Economics 121, 111–141.

Archakov, I., Hansen, P. R., 2021. A new parametrization of correlation matrices. Econometrica 89, 1699–1715.

Archakov, I., Hansen, P. R., Lunde, A., 2020. A multivariate realized GARCH model. Working paper, University of North Carolina, Chapel Hill.

Audrino, F., Knaus, S. D., 2016. Lassoing the HAR model: A model selection perspective on realized volatility dynamics. Econometric Reviews 35, 1485–1521.

Audrino, F., Trojani, F., 2011. A general multivariate threshold GARCH model with dynamic conditional correlations. Journal of Business & Economic Statistics 29, 138–149.

Back, K. E., 2010. Asset pricing and portfolio choice theory. Oxford University Press, Oxford, UK.

Bali, T. G., Beckmeyer, H., Moerke, M., Weigert, F., 2023. Option return predictability with machine learning and big data. Review of Financial Studies, forthcoming.

Bali, T. G., Goyal, A., Huang, D., Jiang, F., Wen, Q., 2022. Predicting corporate bond returns: Merton meets machine learning. Working paper, Georgetown University, University of Lausanne, Swiss Finance Institute, Singapore Management University, and Central University of Finance and Economics.

Barberis, N., Shleifer, A., Wurgler, J., 2005. Comovement. Journal of Financial Economics 75, 283–317.

Barndorff-Nielsen, O. E., Kinnebrock, S., Shephard, N., 2010. Measuring downside risk: Realised semivariance. In T. Bollerslev, J. Russell, and M. Watson, eds., Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle.

Barndorff-Nielsen, O. E., Shephard, N., 2004. Econometric analysis of realized covariation: High-frequency based covariance, regression, and correlation in financial economics. Econometrica 72, 885–925.

Bollerslev, T., 1990. Modeling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. Review of Economics and Statistics 72, 498–505.

Bollerslev, T., 2022. Realized semi(co)variation: Signs that all volatilities are not created equal. Journal of Financial Econometrics 20, 219–252.

Bollerslev, T., Hood, B., Huss, J., Pedersen, L. H., 2018a. Risk everywhere: Modeling and managing volatility. Review of Financial Studies 31, 2729–2773.

Bollerslev, T., Li, J., Patton, A. J., Quaedvlieg, R., 2020a. Realized semicovariances. Econometrica 88, 1515–1551.

Bollerslev, T., Medeiros, M. C., Patton, A. J., Quaedvlieg, R., 2022a. From zero to hero: Realized partial (co)variances. Journal of Econometrics 231, 348–360.

Bollerslev, T., Patton, A. J., Quaedvlieg, R., 2018b. Modeling and forecasting (un)reliable realized covariances for more reliable financial decisions. Journal of Econometrics 207, 71–91.

Bollerslev, T., Patton, A. J., Quaedvlieg, R., 2020b. Multivariate leverage effects and realized semicovariance GARCH models. Journal of Econometrics 217, 411–430.

Bollerslev, T., Patton, A. J., Zhang, H., 2022b. Equity clusters through the lens of realized semicorrelations. Economics Letters 211, 110245.

Bondarenko, O., Bernard, C., 2021. Option-implied dependence and correlation risk premium. Working paper, University of Illinois at Chicago and Vrije Universiteit Brussel.

Boyer, B. H., 2011. Style-related comovement: Fundamentals or labels. Journal of Finance 66, 307–332.

Bucci, A., 2020. Realized volatility forecasting with neural networks. Journal of Financial Econometrics 18, 502–531.

Buraschi, A., Kosowski, R., Trojani, F., 2014. When there is no place to hide: Correlation risk and the cross-section of hedge fund returns. Review of Financial Studies 27, 581–616.

Buss, A., Vilkov, G., 2012. Measuring equity risk with option-implied correlations. Review of Financial Studies 25, 3113–3140.

Cappiello, L., Engle, R. F., Sheppard, K., 2006. Asymmetric dynamics in the correlations of global equity and bond returns. Journal of Financial Econometrics 4, 537–572.

Chan, L. K. C., Karceski, J., Lakonishok, J., 1999. On portfolio optimization: Forecasting covariances and choosing the risk model. Review of Financial Studies 12, 937–974.

Chen, A. Y., Zimmermann, T., 2022. Open source cross-sectional asset pricing. Critical Finance Review 27, 207–264.

Chen, H., Chen, S., Chen, Z., Li, F., 2019. Empirical investigation of an equity pairs trading strategy. Management Science 65, 370–389.

Chen, L., Pelger, M., Zhu, J., 2023. Deep learning in asset pricing. Management Science, forthcoming.

Christensen, K., Siggaard, M., Veliyev, B., 2023. A machine learning approach to volatility forecasting. Journal of Financial Econometrics, forthcoming .

Corsi, F., 2009. A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics 7, 174–196.

Cosemans, M., Frehen, R., Schotman, P. C., Bauer, R., 2016. Estimating security betas using prior information based on firm fundamentals. Review of Financial Studies 29, 1072–1112.

Diebold, F. X., Mariano, R. S., 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13, 253–263.

Driessen, J., Maenhout, P. J., Vilkov, G., 2009. The price of correlation risk: Evidence from equity options. Journal of Finance 64, 1377–1406.

Driessen, J., Maenhout, P. J., Vilkov, G., 2013. Option-implied correlations and the price of correlation risk. Working paper, Tilburg University, INSEAD and Goethe University Frankfurt.

Engle, R., 2002. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. Journal of Business & Economic Statistics 20, 339–550.

Epps, T. W., 1979. Comovements in stock prices in the very short run. Journal of the American Statistical Association 74, 291–298.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56.

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Fama, E. F., French, K. R., 2020. Comparing cross-section and time-series factor models. Review of Financial Studies 33, 1891–1926.

Fan, J., Furger, A., Xiu, D., 2016. Incorporate global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. Journal of Business & Economic Statistics 34, 489–503.

Fleming, J., Kirby, C., Ostdiek, B., 2003. The economic value of volatility timing using "realized" volatility. Journal of Financial Economics 67, 473–509.

Giglio, S., Kelly, B., Xiu, D., 2022. Factor models, machine learning, and asset pricing. Annual Review of Financial Economics 14, 337–368.

Green, J., Hand, J. R. M., Zhang, X. F., 2013. The supraview of return predictive signals. Review of Accounting Studies 18, 692–730.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Review of Financial Studies 33, 2223–2273.

Hameed, A., Morck, R., Shen, J., Yeung, B., 2015. Information, analysts, and stock return comovement. Review of Financial Studies 28, 3153–3187.

Hansen, P. R., Lunde, A., 2005. A realized variance for the whole day based on intermittent high-frequency data. Journal of Financial Econometrics 3, 525–554.

Hansen, P. R., Lunde, A., Voev, V., 2014. Realized beta GARCH: A multivariate GARCH model with realized measures of volatility. Journal of Applied Economertrics 29, 774–799.

Herskovic, B., Kelly, B., Lustig, H., Van Nieuwerburgh, S., 2016. The common factor in idiosyncratic volatility: Quantitative asset pricing implications. Journal of Financial Economics 119, 249–283.

Hoberg, G., Phillips, G., 2010. Product market synergies and competition in mergers and acquisitions: A text-based analysis. Review of Financial Studies 23, 3773–3811.

Hoberg, G., Phillips, G., 2016. Text-based network industries and endogenous product differentiation. Journal of Political Economy 124, 1423–1465.

Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: An investment approach. Review of Financial Studies 28, 650–705.

Israelsen, R. D., 2016. Does common analyst coverage explain excess comovement? Journal of Financial and Quantitative Analysis 51, 1193–1229.

Jagannathan, R., Ma, T., 2003. Risk reduction in large portfolios: Why imposing the wrong constraints helps. Journal of Finance 58, 1651–1683.

Kaniel, R., Lin, Z., Pelger, M., Nieuwerburgh, S., 2022. Machine-learning the skill of mutual fund managers. Working paper, University of Rochester, Stanford University, and Columbia University.

Kelly, B. T., Malamud, S., Zhou, K., 2022. The virtue of complexity in return prediction. Working paper, Yale University and Swiss Finance Institute.

Krishnan, C., Petkova, R., Ritchken, P., 2009. Correlation risk. Journal of Empirical Finance 16, 353–367.

Kumar, A., Lee, C. M., 2006. Retail investor sentiment and return comovements. Journal of Finance 61, 2451–2486.

Li, B., Rossi, A., 2021. Selecting mutual funds from the stocks they hold: a machine learning approach. Working paper, Wuhan University and Georgetown University.

Li, S. Z., Tang, Y., 2022. Automated risk forecasting. Working paper, Rutgers University.

Li, S. Z., Yuan, P., Zhou, G., 2023. Risk momentum: A new class of price patterns. Working paper, Rutgers University, Renmin University of China, and Washington University at St. Louis.

Lintner, J., 1965. The valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets. Review of Economics and Statistics 47, 13–37.

Menzly, L., Ozbas, O., 2010. Market segmentation and cross-predictability of returns. Journal of Finance 65, 1555–1580.

Mueller, P., Stathopoulos, A., Vedolin, A., 2017. International correlation risk. Journal of Financial Economics, 126, 270–299.

Muslu, V., Rebello, M., Xu, Y., 2014. Sell-side analyst research and stock comovement. Journal of Accounting Research 52, 911–954.

Noureldin, D., Shephard, N., Sheppard, K., 2012. Multivariate high-frequency-based volatility (HEAVY) models. Journal of Applied Econometrics 27, 907–933.

Oh, D. H., Patton, A. J., 2016. High-dimensional copula-based distributions with mixed frequency data. Journal of Econometrics 193, 349–366.

Parsons, C. A., Sabbatucci, R., Titman, S., 2020. Geographic lead-lag effects. Review of Financial Studies 33, 4721–4770.

Patton, A. J., Sheppard, K., 2015. Good volatility, bad volatility: Signed jumps and the persistence of volatility. Review of Economics and Statistics 97, 683–697.

Pindyck, R. S., Rotemberg, J. J., 1993. The comovement of stock prices. Quarterly Journal of Economics 108, 1073–1104.

Pirinsky, C., Wang, Q., 2006. Does corporate headquarters location matter for stock returns? Journal of Finance 61, 1991–2015.

Pollet, J. M., Wilson, M., 2010. Average correlation and stock market returns. Journal of Financial Economics 96, 364–380.

Rapach, D. E., Strauss, J. K., Zhou, G., 2013. International stock return predictability: What is the role of the United States? Journal of Finance 68, 1633–1662.

Rapach, D. E., Zhou, G., 2022. Asset pricing: Time-series predictability. Oxford Research Encyclopedia of Economics and Finance .

Roll, R., 1984. A simple implicit measure of the effective bid-ask spread in an efficient market. Journal of Finance 39, 1127–1139.

Sharpe, W. F., 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance 19, 425–442.

Shi, F., Shu, L., Yang, A., He, F., 2020. Improving minimum-variance portfolios by alleviating overdispersion of eigenvalues. Journal of financial and quantitative analysis 55, 2700–2731.

Stambaugh, R. F., Yu, J., Yuan, Y., 2012. The short of it: Investor sentiment and anomalies. Journal of Financial Economics 104, 288–302.

Stambaugh, R. F., Yuan, Y., 2016. Mispricing factors. Review of Financial Studies 30, 1270–1315.

Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society, Series B 58, 267–288.

Tse, Y. K., Tsui, A. K. C., 2002. A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. Journal of Business & Economic Statistics 20, 351–362.

Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21, 1455–1508.