A Survey of Models and Methods Used for Forecasting When Investing In Financial Markets*

Kenwin Maung and Norman R. Swanson

Rutgers University

September 2024

Abstract

The *Makridakis M6 Financial Duathalon* competition builds on prior M-competitions that focus on the properties of point and probabilistic forecasts of random variables by also evaluating investment decisions in financial markets. In particular, the M6 competition evaluates both forecasts and investment outcomes associated with the analysis of a large group of financial time series variables. Given the importance of return and risk forecasting when making investment decisions, a natural question in this context concerns what sorts of methods and models are available for said forecasting, and were used by participants of the competition. In this survey, we discuss such methods and models, with specific focus on the construction of financial time series forecasts using approaches designed for both both discrete and continuous time setups, and using both small and large (high dimensional and/or high frequency) datasets. Examples covered range from simple random walk type models of returns to parametric GARCH and nonparametric integrated volatility methods for forecasting volatility (risk). We also present the results of a novel empirical illustration that underscores the difficulty in forecasting financial returns, even when using so-called big data.

Keywords: Forecasting, Investment, Financial Markets, Big Data, Machine Learning, GARCH Models, Continuous Time Finance Models.

---

* *Corresponding Author:* Norman R. Swanson, Department of Economics, Rutgers University, New Brunswick, NJ, USA, 08901, nswanson@econ.rutgers.edu. Kenwin Maung, Department of Economics, Rutgers University, New Brunswick, NJ, USA, 08901, gm828@economics.rutgers.edu. This paper has been prepared for the special issue on the Makridakis M6 Forecasting and Investment Competition. The authors are grateful to the organizers of the competition, the organizers and participants of the associated conference in New York City in December 2023, and the editors of the special issue for guidance received during the writing of the paper. The authors are also grateful to Kaiwen Qiu for excellent research assistance, and to John Chao, John Landon-Lane, Yuan Liao, and Minchul Shin for useful comments.

1. Introduction

In the 25 years since the release of the acclaimed book entitled *A Non-Random Walk Down Wall Street* by Lo and MacKinlay (1999), the finance and forecasting fields have benefited greatly from the development of an impressive array of new methods, models, and tools designed to aid in the construction of forecasts useful for investing in the financial markets. These developments have not only been spurred on by the convincing arguments made by Lo, MacKinlay, and many others that not all markets evolve over time like a "drunk man walking", but also by the tremendous increase in the amount of data available with which to model and forecast financial time series. In much of this survey, we summarize a number of advances made over this time period for modelling and forecasting using datasets that are high frequency and high dimensional datasets (e.g. "big data") as well as using more traditional datasets that may contain relatively few variables and/or observations. For the latter type of dataset, there are a number of often rather simple methods that have been used for modelling and forecasting financial variables for many decades. We also give shrift to such methods, but our focus is primarily on more recent methods.

Why is forecasting so important when discussing investing in finance? The reason for this is that much of the investment world utilizes modern quantitative investment management, and quantitative investment management equates with data-driven decision making in the markets. For discussions of quantitative equity portfolio management and quantitative portfolio optimization see Rasmussen (2003) and Chincarini and Kim (2006). This sort of management and optimization makes use of a variety of models and methods to characterize equities, fixed income products, currencies, commodities, and a whole host of other structured financial instruments. These models and methods are designed to solve problems related to asset pricing and hedging, risk analytics, and portfolio optimization, and in contexts where the goal is to allocate assets and optimize portfolios, forecasting key quantities like risk and return is important.

As discussed above, in this survey we focus on models and methods used for forecasting. These include simple approaches based on the use of judgement and surveys, and well as simple models such as historical averages, random walks, and autoregressive regressions. Another simple approach that we mention is that based on so-called technical analysis, which uses charts to assess market conditions. The relative trade-offs associated with using models of judgement versus quantitative models is discussed in Makridakis et al. (2024).

Needless to say, though, much interest in recent years has centered on the development of more complex methods and models that fully take into account to availability of large high dimensional datasets that

may contain ultra-high frequency observations such as stock price data on every trade made on a particular company. In addition to being used to construct high frequency measures of returns, such datasets are used, for example, to construct estimates of daily (unobserved) volatility by summing up intra-daily volatility (risk) measurements made every 1, 2, or 5 minutes, for example. Models and methods designed to be implemented in these sorts of data-rich environments that are discussed in the sequel include two types - those based on "discrete modeling" approaches and those based on "continuous modelling" approaches. Broadly speaking, "discrete modelling approaches" are those designed for data that are assumed to be discrete, in the sense that the time interval between observations is not assumed to go to zero when deriving asymptotic properties of the models. Alternatively, 'continuous modelling" approaches are those designed for data that are assumed to be continuous, in the sense that the time interval between observations is assumed to go to zero when deriving asymptotic properties of the models.

Some of the 'discrete modeling' approaches that we discuss include time varying parameter models and big data or machine learning methods based on as factor models, neural networks, random forests, the elastic net, and the least absolute shrinkage and selection operator. Some of the "continuous modeling" approaches that we discuss include realized volatility and other jump and noise robust measures of integrated volatility. These measures are nonparametric, as they do not require the specification of an underlying model. We also discuss various continuous time model of return and (stochastic) volatility including geometric Brownian motion, the classical Cox-Ingersoll-Ross model (Cox et al., 1985), and a variety of stochastic volatility models.

The rest of the paper is organized as follows. We begin with a very brief discussion of the importance of forecasting when carrying out financial investment, in Section 2. We then discuss methods and models for forecasting returns in Section 3, and approaches to forecasting volatility in Section 4. Methods for evaluating forecasts and investment performance are discussed in Section 5, and a small empirical illustration is given in Section 6. Finally, concluding remarks are gathered in Section 7.

2. The Role of Forecasting in Finance

Financial forecasting is instrumental in constructing optimal investment portfolios and helps to inform the pricing of individual assets. Modern portfolio theory (MPT), due to Markowitz (1952), and the Capital Asset Pricing Model (CAPM) are two fundamental frameworks that use expected returns and risks as inputs, helping market participants make investment decisions. Forecasting in finance is important, for example, because estimates of expected returns and future risks are frequently derived from forecasts.

MPT offers a framework for building and selecting portfolios depending on an investor's risk tolerance

and the expected performance of assets. The MPT investment process takes estimates of individual asset returns, volatilities, and their correlations as given, along with constraints on investment choices (e.g. turnover constraints), to perform a mean-variance optimization (i.e. selecting portfolio weights) that results in a portfolio yielding either: (i) the maximum possible expected return for a given level of risk or (ii) the smallest risk possible for a given level of expected returns (see Fabozzi et al. (2002)). It is worth stressing that the primitives of this investment process (e.g. the expected returns and volatilities of individual assets) are unobservable objects and thus have to be estimated beforehand. Common estimators of expected returns and volatility are the historical mean and standard deviation of returns, respectively, as discussed in Miccolis and Goodman (2012). However, estimates based solely on historical returns data can aggregate information from very different economic and market regimes, potentially resulting in forecasts that do not accurately reflect any specific environment. A natural solution to this problem is to incorporate external predictor information, such as macroeconomic indicators and market conditions, into models for expected returns and volatilities. Given recent increases in the scope and availability of data, integrating external predictors in these models has become more common, more feasible, and more valuable. Additionally, data-rich environments have enabled and served as motivation for the use of machine learning algorithms that can process vast quantities of data and adapt to changing market dynamics more effectively than many traditional 'data-poor' models. We return to this discussion of how investors can leverage these advanced techniques to improve their predictions of future returns and volatilities in Sections 3 and 4.

Asset pricing theories offer an alternative method for calculating expected returns of individual assets or classes. For example, the CAPM argues that the excess return of an asset (expected returns minus the risk-free rate) is proportional to the equity premium (expected market return minus the risk-free rate). The proportionality constant is called 'beta' and is a measure of an asset's risk relative to that of the market. One can use this model to estimate expected returns via a two-pass regression, as discussed in Fama and MacBeth (1973) (see also Bartholdy and Peare (2003)), by using the historical average of the equity premium as a proxy for the expected equity premium. Alternatively, one can form a forecast of the expected equity premium to be subsequently used in the CAPM. The methods and challenges related to the forecasting of the equity premium are themselves the subject of many comprehensive studies (see Goyal and Welch (2003), Rapach and Zhou (2013), and Goyal et al. (2023)). We return to these issues in the following Section.

Ultimately, regardless of one's investment framework, quantitative asset management fundamentally

requires expectations of future returns and risks. Asset managers must anticipate future market volatility, economic conditions, and crucially, how these factors impact investments. Moreover, effective investment and risk management strategies, such as diversification or hedging, depend on how expectations and forecasts of return and risk evolve over time. With this in mind, we review a range of forecasting approaches beginning with judgemental forecasting and simple time-series models, and moving on to more data-intensive machine learning techniques suitable in the current age of "big data".

3. Methods and Models for Forecasting Returns

As prefaced, returns forecasting is valuable on many fronts. It is fundamental for the efficient allocation of capital by portfolio managers and other practitioners who use real-time forecasts in investment decisions, such as adjusting asset weights in portfolios to optimize performance. Corporations also rely on forecasts when making strategic decisions involving capital investments, mergers, and acquisitions. Additionally, returns forecasting is crucial for validating and testing asset pricing models, which are used to determine the intrinsic value of financial instruments and to help identify mis-priced assets.

A natural question to ask is whether it is possible to accurately forecast stock returns. Indeed, a central question that features in many of the proposed forecasting methods and models is whether they are able to outperform the forecasts implied by the historical averages and standard deviations of returns, which often turns out to be an incredibly challenging task. In the context of returns, this is significant because the historical average forecast is consistent with the idea that (log) stock prices follow a random walk (with drift), and are therefore unpredictable. In this section, we discuss methods and models of returns that are currently used in quantitative analysis by a whole host of active quantitative analysts working in the financial markets. In the following section, we do the same for volatility (risk).

3.1. Simple Methods and Models of Returns

3.1.1. Judgmental Forecasts

Forecasters often consult their own experiences and opinions when forming forecasts of returns. Such judgments might be the only information used to guide quantitative model based investing, for example, and might be purely intuitive, or instead might be based off merging intuition with model based findings. Needless to say, the same goes when forecasting volatility - judgment forecasts are often utilized by practitioners. For example, judgment can help to: (i) define the space of candidate forecasting models for consideration and aid in the eventual selection of a model, (ii) determine how a model can be used to produce forecasts (e.g. should a rolling or expanding window be used when estimating a model using

historical data?), and (iii) inform subjective adjustments of model-implied forecasts to cohere with one's prior experience and/or intuition (Petropoulos et al., 2018). In practice, reliance on judgment alongside quantitative models is commonly observed, as evidenced by the approach used by many participants of the Survey of Professional Forecasters, conducted separately in the US (Federal Reserve Bank of Philadelphia) and in Europe (European Central Bank), who often report using judgment along with reduced-form and structural models, to produce forecasts (Clements et al., 2023; de Vincent-Humphreys et al., 2019; Stark, 2013).

Many authors have found that judgmental forecasting can produce rather accurate predictions that are on par with or outperform the best purely quantitative models, particularly if the forecaster has good domain knowledge (Lawrence et al., 2006). However, this is not guaranteed. In the recent M6 forecasting competition (see Makridakis et al. (2024)), participants were tasked with predicting the performance of 100 publicly traded assets over the course of an entire calendar year. Submitting teams could declare whether they were using 'pure judgment', 'judgment-informed' or 'data-driven' approaches. Submissions that relied heavily on judgment were generally inferior when compared to data-driven methods although a minority of judgment-informed submissions did outperform (Makridakis et al., 2024). Furthermore, only a small number of submissions (8.8%) self-identified as having used 'pure judgment' or 'judgment-informed' approaches while the majority (68.4%) relied on time series or machine learning models of the variety surveyed in this paper. Nonetheless, it is conceivable that many 'data-driven' teams exercised judgment in forming, selecting, and parameterizing their quantitative models, although the extent to which this sort of judgment affected or contributed to forecasting performance was unobserved.

A specific application of subjective forecasting in returns prediction is technical analysis, also known as the *chartist* approach. Typically, this involves identifying regularities in time series graphs of asset price and volume data, with the expectation that these patterns can inform future movements. Lo et al. (2000) argue that this is highly subjective or judgmental because the identification of geometric shapes or patterns in a graph is often in the 'eye of the beholder', and further point out that technical analysis has been compared to the unscientific disciplines of astrology and alchemy. These authors attempt to suppress this subjectivity by introducing a systematic approach which attempts to: (i) define key patterns in terms of geometric properties (e.g. using a sequence of local extrema), (ii) construct nonparametric estimators for asset prices, and (iii) analyze these estimators for the occurrence of each geometric pattern. Using this procedure, they find that several technical patterns can be predictive of future price movement, particularly for NASDAQ stocks. Nonetheless, this approach still requires the development of a lexicon of

pre-defined patterns that are believed to be predictive of future trends. A more recent approach by Jiang et al. (2023) dispenses with this requirement by learning predictive patterns from the stock price images via the use of convolutional neural networks, and subsequently producing out-of-sample predictions.

A related approach involves using *sentiment analysis* based on machine learning techniques to explicitly capture subjective opinions expressed in news articles, announcements, or social media for use in financial forecasting. Sentiment is associated with the notion of emotional-based evaluation, which may in turn influence judgment via several channels, as discussed in Kabiri et al. (2023). The idea is that an agent's sentiment towards financial markets may be useful for forecasting. Frydman et al. (2021) supports this notion by providing evidence that when market sentiment is optimistic and good news on dividends and interest rates arrive, this news has significant positive impact on survey participants' forecasts of future returns. Nonetheless, it has to be acknowledged that sentiment-driven stock price movements can occur in the absence of changes to fundamentals such as news on dividends and interest rates. A case in point is the 'meme' stocks that experienced surges in prices during the early 2020s (for example, at one point there was a price increase of over 700% for GameStop) because of the 'to the moon' movement coordinated by the subreddit r/WallStreetBets on Reddit. This type of movement can be attributed to herding behavior or noise traders (Long et al., 2023).

3.1.2. Random Walks and the Historical Averages

We begin with a parsimonious model-based approach for our forecasting process. Consider a simple linear model for one-period ahead returns forecasting:

$$r_{t+1} = \beta_0 + x_t'\beta + u_{t+1}, \tag{1}$$

where r_{t+1} is the change in the aggregate market price from period t to $t+1$ in excess of the risk-free rate[1] (i.e. returns from a risky asset minus the risk-free rate), and u_{t+1} is some forecast error with zero-mean. In the literature, common variables used for the aggregate market return are the log difference of the S&P500 index, and the CRSP value-weighted equity returns. Here, x_t is a vector of predictor variables[2].

The simplest 'forecasting' model to consider in the framework of (1) is the random walk model, which

[1]This is often referred to as excess returns or the equity premium in the case of stock returns. Note that the techniques that we discuss here can also also be applied to the returns of individual assets or classes.

[2]Given that we are interested in forecasting, our analysis will focus on out-of-sample predictability of excess returns.

can be obtained by setting $\beta_0 = 0$ and $\beta = 0$. For simplicity, ignore the risk-free rate. This yields:

$$\ln P_{t+1} - \ln P_t = r_{t+1} = u_{t+1},$$

where P_t is the price of an asset at time t. This model corresponds to the view that stock returns cannot be predicted in a consistent fashion as the difference in (log) prices is essentially an unforecastable random disturbance. The Efficient Market Hypothesis can be used to motivate such a specification. In an efficient market, the price of an asset should reflect its intrinsic value. When market participants, who are assumed to have a noisy estimate of its value, trade en masse, the market price of the asset will fluctuate randomly around this instrinsic value (Fama, 1965, 1995). Under the random walk model, our best guess of next period's asset price is simply just the realized price that we observe today.

When $\beta_0 \neq 0$, we obtain a random walk with drift model. One way to motivate this model, given the above explanation, is that the intrinsic value of assets might vary over time due to but not limited to news of research and development or due to changes in management (Fama, 1965). Given this specification, our forecast, $\hat{r}_{t+1}$ of stock returns at $t+1$ is the historical average at time t (the mean of all returns up to time t), also known as the prevailing mean:

$$\hat{r}_{t+1} = \sum_{i=1}^{t} r_i.$$

To see why this works, observe that the model is:

$$r_{t+1} = \beta_0 + u_{t+1}.$$

Let $E_t r_{t+1} = E(r_{t+1}|F_t)$, where F_t is the filtration of all information through time t (or the information set available at time t). It follows that $E_t r_{t+1} = \beta_0$, whose consistent estimator would be the historical average. This model also implies that a natural forecast of $\ln P_{t+1}$ is simply $\ln P_{t+1} + \beta_0$, so that prices are assuming to 'drift' upwards over time, on average, with the average drift being equal to β_0.

When forecasting excess returns out-of-sample, Goyal and Welch (2003, 2008) find that the prevailing mean is an incredibly difficult benchmark to outperform, and many forecasts conditioned on additional predictor variables fail to beat it in a statistically significant manner. This has led to the development of many sophisticated models designed to deliver forecasts superior to those implied by the above modeling approach, some of which are discussed in the remainder of this survey.

3.1.3. Predictive Regressions

The set-up in (1) allows us to include additional predictors in x_t. Goyal and Welch (2008) provides a comprehensive review of key predictors that have been argued to be useful in forecasting excess returns.

These authors construct out-of-sample forecasts using predictors which include financial ratios (e.g. dividend price ratios and book-to-market ratios), stock characteristics (e.g. stock variance and cross-sectional premia), and macroeconomic variables (e.g. inflation and investment to capital ratios). Their approach involves estimating univariate regressions of excess returns on each predictor (i.e. estimating (1), where x_t is univariate). The out-of-sample performance of each of the forecasts based on these models is then compared with forecasts implied by the historical average.[3] In their full sample analysis, only 2 out of the 17 predictors (equity issuing activity (Baker and Wurgler, 2000) and the consumption, wealth, and income ratio (Lettau and Ludvigson, 2001)) performed consistently better than the historical average in out-of-sample forecasting. In response to this result, Campbell and Thompson (2008) show that out-of-sample predictability can be improved for many of the original predictors by imposing theoretically motivated restrictions on the forecasting model in (1). They apply 2 restrictions either sequentially or jointly, the first being that the regression coefficient, β, has to have the theoretically expected sign, and the second being that the predicted value of the excess return is non-negative. Pettenuzzo et al. (2014) propose an additional restriction on the conditional Sharpe ratio such that it is bounded by 0 and an upper bound, so that the price of risk is not too high. Under these proposed restrictions, the above authors report additional significant statistical and economic gains relative to both the unconstrained case and to the restricted case suggested by Campbell and Thompson (2008).

In a recent update to their original paper, Goyal et al. (2023) repeat their analysis with updated data (up to 2022) and 29 new predictors that have been recently introduced in the literature. These predictors can be broadly related and grouped into 6 categories: macroeconomic, sentiment, stock variance, stock cross-section, other stock market characteristic, and commodities. They also reconsidered the performance of the original 17 predictors. The results indicate that the following predictors performed favorably out-of-sample: (i) a principal component of 14 technical indicators (Neely et al., 2014), (ii) aggregate short interest in the stock market (Rapach et al., 2016), (iii) aggregate accruals (Hirshleifer et al., 2009), (iv) the fourth-quarter growth in personal consumption expenditures (Møller and Rangvid, 2015), (v) treasury bill rate (Campbell, 1987), (vi) equity issuing activity (Baker and Wurgler, 2000), (vii) investment-capital

[3]For example, compare the ratios of mean squared errors (MSE): $\bar{R} = 1 - MSE_P/MSE_H$, where MSE_P is the MSE of the forecast conditioned on a predictor and MSE_H is the MSE of the historical average forecast. Here, a negative $\bar{R}$ implies that the historical average performed better, and vice-versa. Subsequently, the MSE-F test of McCracken (2007) can be used to check the statistical significance of the ratio.

ratio (Cochrane, 1991)[4]. The first 4 predictors were newly introduced while the latter 3 were present in their original examination. One key finding is that the superior predictive performance of these predictors frequently varies over time. Specifically, the authors point out that the aggregate accrual predictor derives its good performance almost exclusively from the dot-com bubble episode and has relatively modest performance elsewhere. We will return to this idea of time-varying predictability in a subsequent section of this paper.

3.2. Advanced Methods and Models of Returns

3.2.1. Big Data and Predictor Selection - Machine Learning Part I

The set-up in (1) allows for multiple predictors in x_t, however, we have only discussed univariate forecasting models thus far. Goyal and Welch (2008) specifies a 'kitchen sink' regression which includes all predictors in one regression but the out-of-sample forecasting performance of their model is particularly poor. Reasons for this could be the increased estimation noise introduced when estimating so many parameters or possibly an issue with in-sample overfitting. In this section, we discuss econometric and machine learning methods that are used to counteract such issues and that enable us to effectively forecast excess returns using a large number of predictors.

First, let the number of predictors in x_t be P, and the sample size used to estimate (1) be T. We consider two different cases: (i) $P > T$ and (ii) $T \geq P$. The first scenario often occurs when we use a large number of predictors, and is an common situation given recent advances in data collection and reporting technologies (e.g. think of high dimensional and high frequency financial datasets). For example, Dong et al. (2022) use 100 long-short anomalies from the cross-sectional predictability literature to forecast aggregate market return.

When we have a high-dimensional set of predictors, the predictive regression in (1) cannot be estimated via ordinary least squares[5] (OLS). A common solution is to assume a framework of sparsity.[6] Here, we

[4]Here, the criteria for 'good' performance is that the predictor-based forecasts have to have both statistically significant in-sample predictability and statistically superior forecast performance, compared to the benchmark historical average. Furthermore, they are subject to the restriction of not predicting a negative excess return (Campbell and Thompson, 2008). Depending on the evaluation sample period used, the authors remark that 3 more predictors also displayed good forecasting performance.

[5]More precisely, there is no unique solution to the least squares problem of $P > T$.

[6]Another approach assumes that all coefficients in a regression model are non-zero, but uses so-called ridge-regression to estiamte the model.

assume that a large number of regression coefficients are zero. This means that our high-dimensional regression is actually a low-dimensional model because it effectively has only a few regressors. Intuitively, this approach is consistent with the idea that not all predictors are relevant or useful in forecasting returns. However, the key problem is that we do not know, *a priori*, which predictors are relevant.

This problem can be solved with regularization or penalized regression (e.g. ridge regression is a form of penalized regression). The poster child of this approach is the least absolute shrinkage and selection operator (LASSO). Implementing the LASSO involves solving the following penalized least squares regression problem associated with equation (1):

$$\min_{\theta=(\beta_0,\beta')'} \sum_{t=1}^{T} \left(r_{t+1} - \beta_0 - x_t'\beta\right)^2 + p_\lambda(\theta) \tag{2}$$

where

$$p_\lambda(\theta) = \lambda \sum_{i=0}^{P} |\beta_i|, \tag{3}$$

and λ is called the regularization or tuning parameter. Depending on the value of λ, the LASSO estimator can yield sparse estimates of the regression coefficient vector, θ. In general, when λ is large, the solution vector tends to be more sparse because the large positive penalty associated with the 'penalty term' involving λ can be offset by setting some of the regression coefficients, $|\beta_i|$, to 0 (see Hastie et al. (2015)). Often, λ is chosen via a data-driven method such as cross-validation, so that the discovery of which predictors are relevant is an automatic process.[7]

Even when $T \geq P$, so that OLS is feasible, regularization can improve estimation accuracy of the predictive regression (i.e. reduce the MSE).[8] As demonstrated above, regularization can select a smaller model and reduce model complexity, which translates into a lower estimator variance. Nonetheless, when we introduce penalization, we deviate from the OLS estimator which is the best linear unbiased estimator

[7]There are crucial issues regarding the selection properties of the LASSO estimator. A sufficient condition for correct (asymptotic) selection of the LASSO requires relevant predictors to be orthogonal to irrelevant predictors, which is highly unlikely to hold when it comes to financial and macroeconomic data. Hence, in a finite sample, it is likely that it will not correctly select predictors. Newer penalty terms that require weaker conditions have been suggested. Examples are the adaptive LASSO (Zou, 2006) and the smoothly clipped absolute deviation penalty (Fan and Li, 2001). Both approaches allow the penalty to differ across the variables, while in the original LASSO, the same penalty is applied to all the predictors. Still, all versions of the LASSO are approximations, and it has been found that the LASSO still performs very well in many empirical applications.

[8]The expected squared deviation of the estimator from its true value is $E\left[(\hat{\theta} - \theta)^2\right]$. This can be shown to be equivalent to the squared bias plus the estimator variance: $[E(\hat{\theta}) - \theta]^2 + Var(\hat{\theta})$.

and thus introduce bias when using the LASSO or variants of it. However, if this additional bias is smaller than the reduction in variance, the MSE of our estimation improves. Lee et al. (2022) revisit the 'kitchen sink' regression with 12 predictors as in Goyal and Welch (2008) and show that the (adaptive) LASSO can improve forecasting performance of market returns when compared to using OLS or simply the historical average of returns.

Advanced variants of the LASSO are often used in the literature due to limitations in the original approach discussed above. Additionally, a key issue is that even if a group of predictors are relevant (i.e. have non-zero coefficients), if they are highly correlated with one another, the LASSO tends to select only one representative out of that group. The elastic net (Zou and Hastie, 2005), which uses a penalty that is a convex combination of the LASSO (ℓ_1) and the ridge penalties (ℓ_2) was introduced to deal with this issue. The penalty is:

$$p_\lambda(\theta) = \lambda \left[(1-\alpha) \sum_{i=0}^{P} \beta_i^2 + \alpha \sum_{i=0}^{P} |\beta_i| \right], \tag{4}$$

and $\alpha \in [0,1]$ determines how much weight is given to either the LASSO or ridge penalty terms (the LASSO penalty term is the one involving absolute values and the ridge penalty involves summing squares of coefficients). For example, when $\alpha = 1$, the elastic net penalty reduces to the LASSO penalty. Rapach et al. (2013) consider a predictive regression of US market returns on lagged market returns and the returns of 10 other international markets, along with predictors such as the t-bill rate and the dividend yield ratio, and estimate a regression using the (adaptive) elastic net. Similarly, Dong et al. (2022) use the elastic net (among other approaches) to estimate a regression with 100 cross-sectional anomalies as predictors and show that their regression model can generate both statistically and economically significant improvements when forecasting out-of-sample.

3.2.2. Factor Models - Machine Learning Part II

Instead of selecting relevant predictors from a high-dimensional set of variables, the factor model framework posits that the predictors can all be explained by a small number of latent common components or factors. In this type of machine learning, estimates of the (unobserved) factors can be used in the predictive regression in place of the original set of individual predictors. A standard model in this framework is called the *dynamic factor model* and is specified as follows:

$$x_{it} = \lambda_i f_t + \varepsilon_{it}, \tag{5}$$

$$f_t = A(L) f_{t-1} + e_t, \tag{6}$$

$$r_{t+1} = f_t' \beta + w_t' \delta + u_{t+1}. \tag{7}$$

Here, equation (5) posits that the i-th predictor in x_t can be linearly explained by a $k \times 1$ vector of common factors, f_t, that are unobserved. Additionally, ε_{it}, e_t, and u_t are stochastic disturbance terms. The dependence in this model on the factors is determined by the $1 \times k$ vector of factor loadings, λ_i. The factors themselves can follow a VAR(p) process (see (6)) as characterized by the lag polynomial matrix $A(L)$. Assuming that we can identify and estimate the factors from the observed predictors, x_{it}, we can further use them in the predictive regression in (7), additionally accommodating other observed variables in w_t (such as lags of r_t and lags of other key predictor variables).

As this model contains several unobserved objects, we first consider the issue of identification. We stack λ_i over all predictors, for $i = 1, ..., P$, into the $P \times k$ matrix of factor loadings, Λ, and rewrite (5) as:

$$x_t = \Lambda f_t + \varepsilon_t, \tag{8}$$

where ε_t is the stacked vector of ε_{it}. Note that for an invertible matrix, A, of conformable dimension, this model is observationally equivalent to: $x_t = (\Lambda A)(A^{-1} f_t) + \varepsilon_t \equiv \Lambda^* f_t^* + \varepsilon_t$. This problem arises because both Λ and f_t are latent. To deal with this identification issue, several normalizations have been introduced. We focus on the so-called principal components analysis (PCA) normalization. This involves choosing one of the following 2 sets of normalizations: (i) $F'F/T = I_k$ and $\Lambda'\Lambda$ is a diagonal matrix, where F is the stacked $T \times k$ matrix of factors for the whole sample period; or (ii) $\Lambda'\Lambda/P = I_k$ and $F'F$ is diagonal.[9] Given either of the restrictions, we can uniquely pin down an estimate of the latent factors and loadings.

Given that we are interested in extracting factors for use in predictive regressions, we focus on the nonparametric approach and do not specify the transition in (6).[10] Stock and Watson (2002a) show that we can estimate the factors and loadings in the above model by solving the following minimization problem:

$$\min_{(F,\Lambda)} Tr\left\{(X - F\Lambda')(X - F\Lambda')'\right\}, \tag{9}$$

subject to either normalization scheme described above, where $Tr\{\cdot\}$ refers to the trace operator, and X is a $T \times P$ matrix of x_t stacked over time. They note that this optimization is equivalent to the principal

[9] There are several alternative ways to write these normalization restrictions. For example, they may only hold asymptotically if one views the factors and loadings as random variables. For a summary of more identifying restrictions, see table 1 of Bai and Li (2012).

[10] In other words, we allow the dynamics of the factors to be arbitrary (to an extent). If one is interested in recovering the parameters in (6), a state space approach can be considered (see Stock and Watson, 2016).

components analysis problem. For exposition, assume that we have used normalization (ii), and the PCA solution is obtained by setting $\hat{\Lambda}$ to be the first k eigenvectors of $X'X$ that correspond to the k largest eigenvalues of this matrix, and $\hat{F} = X\hat{\Lambda}/P$ is just the resulting least squares estimate. The situation can be symmetrically derived if we have used normalization (i) instead. In this approach, PCA estimates of the factors are consistent for the space spanned by the factors and can subsequently be used in the predictive regression in (7) without having to address the generated regressor problem associated with the fact that the explanatory variables in the predictive regression (i.e. the factors) are themselves estimated and not directly observed. Ludvigson and Ng (2007) estimate factors from 209 macroeconomic indicators and 172 financial variables using principal components analysis, and use them in predictive regressions. These factors exhibit superior performance when forecasting one-quarter ahead excess returns. Çakmaklı and van Dijk (2016) perform a similar analysis with macroeconomic data at the monthly frequency and observe similar results. Dong et al. (2022) extracts the first principal component from their dataset of 100 anomaly predictors and show that it performs comparably well with their elastic net forecast.

The PCA approach can be used even when the number of predictors is small. For example, Neely et al. (2014) extract principal components from 14 macroeconomic variables and 14 technical indicators and show that forecasts of excess return based on these factors perform the best out of all competing models and even outperform the historical average. The market-wide sentiment index of Baker and Wurgler (2006, 2007) is the first principal component of 6 variables that they argue proxies for investor sentiment (for e.g. NYSE share turnover and the dividend premium). However, several studies including Arif and Lee (2014) and Huang et al. (2015) show that this index has little to no predictive power for future aggregate stock returns.[11]

One potential shortcoming in using PCA for forecasting is that the resulting common components are constructed to capture the maximum variation common to all of the predictors. The estimation procedure does not take into account the forecast target when forming the common factors. In other words, it may be the case that some of the factors that capture the maximum variation among all predictors are not best suited for forecasting a particular variable. To see this, we can split the factors in (5) into two parts:

$$x_{it} = \lambda_i^R f_t^R + \lambda_i^E E_t + \varepsilon_{it}, \tag{10}$$

where f_t^R is the part of the common component that is most relevant to forecasting returns while E_t

[11]It should be noted that the authors of the original index did not propose using it for forecasting aggregate returns, but instead proposed using it for cross-sectional asset pricing.

describes the common variation in the predictors irrelevant to returns forecasting. The relevant factors will now be used in the forecasting step:

$$r_{t+1} = f_t^{R\prime}\beta^R + w_t'\delta + u_{t+1}. \tag{11}$$

Kelly and Pruitt (2013, 2015) show that improvements to stock return forecasting can be obtained by focusing on the common components of the predictors that are most relevant to the forecast target. This begs the question of how we can obtain f_t^R in the first place given that PCA is unable to differentiate between f_t^R and E_t. Here, we use the method of Partial Least Squares (PLS) first introduced by Wold (1966) and applied to stock return forecasting by Kelly and Pruitt (2013). This estimation procedure is executed in two steps. We begin with a time-series regression of the individual predictors on the one-period ahead excess returns:

$$x_{it} = \theta_{0i} + \theta_{1i} r_{t+1} + \xi_{it},$$

for each given i. From (11), we have that r_{t+1} is driven by f_t^R, which is one component of x_{it}, hence θ_{1i} approximately captures how the individual predictors depend on the common factors that are relevant to returns forecasting. In other words, θ_{1i} approximates λ_i^R. Recall that we are interested in estimating f_t^R, which is easily obtained via a cross-sectional least squares regression of (10) if we knew what λ_i^R is. Although we do not observe λ_i^R, we have, from the first step, a rough estimate of it in the form of $\hat{\theta}_{1i}$ for all i and thus we employ it in the following cross-sectional regression:

$$x_{it} = a_t + f_t^{R,PLS}\hat{\theta}_{1i} + \tilde{\xi}_{it}.$$

Repeating this regression for all t results in a time-series of $\{f_t^{R,PLS}\}$ which we regard as estimates of $\{f_t^R\}$. Groen and Kapetanios (2016) prove the consistency of the PLS estimator in this context while Kelly and Pruitt (2015) provided asymptotic inference results.

Another approach to the above problem is to consistently select the variables that are relevant for forecasting the target, prior to the construction of consistent factor estimates. Factors estimated in this way are efficient, in the sense that irrelevant variables are excluded when the factors are estimated. One can then proceed to specify a predictive regression by selecting those factors that are relevant for predicting returns via use of a variable selection procedure based on predictive experiments using a training dataset, for example. This method is discussed in Chao et al. (2024).

3.2.3. Time-varying Predictive Regressions

Aggregate stock returns depend on the state variables of the economy, which in turn can vary over the business cycle.[12] Pesaran and Timmermann (1995) and Timmermann (2008) provide early evidence that the out-of-sample predictive power of market return predictors tends to change over time and that predictability is particularly low when the markets are calm but is greatly elevated during volatile periods. In other words, predictability is local to short periods or intervals (later termed 'pockets') of time. It is difficult to see how a static predictive regression as in (1) would be able to match such changes in predictability over time. Hence, we shall discuss the following time-varying predictive regression:

$$r_{t+1} = \beta_{0t} + x_t'\beta_t + u_{t+1}, \quad Var(u_{t+1}) \equiv \sigma^2_{t+1}. \tag{12}$$

A simple parametric model to capture the time variation of β in this model is the following specification:

$$\beta_{0t} = \beta_0(s_t), \quad \beta_t = \beta(s_t), \quad \sigma^2_t = \sigma^2(s_t) \tag{13}$$

where $s_t \in \{1, ..., M\}$ represents a latent state of the economy at time t and M is the maximum number of states. The simplest (and probably most intuitive) example is when $M = 2$, and we interpret the states as being representative of either economic recession and expansion. Another possible interpretation is that of a bull or bear market. The latent states are often assumed to follow a Markov-chain, which implies that (12)-(13) is a Markov-switching model. This model can be estimated by maximum likelihood, often with the EM algorithm, and the latent states can be obtained by filtering and smoothing (see Hamilton, 2016, for modeling and estimation details). Henkel et al. (2011) estimate a Markov-switching vector autoregression (VAR) with two states using the endogenous variables: (i) short-horizon aggregate excess returns, (ii) dividend yield, (iii) short-term interest rate, (iv) term spread, and (v) the default rate. The equation of interest in this VAR is the first equation as it relates excess returns to lagged returns and the other predictors (this is similar to the specification in (12)). They interpret the two estimated latent states as corresponding to recessions and expansions. They find that the R^2 for the first equation (the equation that corresponds to the excess return) is much higher during recessions, which they argue implies greater return predictability (i.e. better model fit) during economic downturns. It is important to remark that their work was done by estimating their model in-sample (i.e. results were obtained using the full sample data that they had available), although the authors do provide some out-of-sample evidence that

[12] See Rapach and Zhou (2013) for a detailed explanation on how return predictability depends on time-varying aggregate risk.

the Markov-switching VAR can outperform the historical average but only during recessions, which is consistent with the paper's message.

The parametric assumption on the evolution of the model parameters in (13) might be too restrictive. One could instead allow $\beta_{0t}, \beta_t, \sigma_t^2$ to be continuous functions of time. Dangl and Halling (2012) consider the following state space model for their predictive regressions:

$$r_{t+1} = z_t'\theta_t + u_{t+1}, \quad u_{t+1} \sim N(0, V)$$
$$\theta_t = \theta_{t+1} + e_t, \quad e_t \sim N(0, W_t),$$

where $z_t = (1, g_t)'$, and g_t here is a univariate predictor. The authors use 13 predictors that were studied by Goyal and Welch (2008) in their univariate time-varying predictive regressions. The model is estimated via Bayesian methods and the analysis is entirely out-of-sample. The key findings are that excess return forecasts from this time-varying model can statistically outperform the historical average and that return predictability is similarly countercyclical. Although they do remark that predictability during expansions is not as low as was previously suggested by Henkel et al. (2011).

Nonetheless, the random walk restriction in the coefficient transition could be problematic as that might lead to non-stationarity in the returns.[13] To ensure a more robust analysis, one could specify that θ_t is an unknown but smooth function of time: $\theta_t = \theta(t)$, and estimate the model with nonparametric techniques.[14] This is the approach taken by Farmer et al. (2023), where the estimator of θ_t is the Nadaraya-Watson local constant estimator:[15]

$$\hat{\theta}_t = \text{argmin}_{\theta_0} \frac{1}{T} \sum_{s=1}^{t-1} (r_{s+1} - z_s'\theta_0) k_{st}, \tag{14}$$

where

$$k_{st} = \frac{1}{h} K\left(\frac{s-t}{Th}\right),$$

$K(\cdot)$ is a kernel function, and h is a positive bandwidth. The authors use the (one-sided) Epanechnikov kernel: $K(u) = 3/2(1-u^2)$ for $-1 < u < 0$, and 0 otherwise. Note that the estimation of the time-varying

[13]The authors do consider an autoregressive model in their appendix, but show that the random walk model outperforms in terms of forecasting.

[14]To be precise, for the consistency of nonparametric estimators, particularly a kernel estimator of θ_t, we have to assume that $\theta_t = \theta(t/T)$, so that it is a function of *standardized* time. Then, the estimator can be shown to be consistent under an infill asymptotic scheme, as $T \to \infty$ (Robinson, 1989; Cai, 2007).

[15]Nonparametric kernel estimators are not commonly used for financial forecasting, but are more commonly used in the macroeconomic forecasting literature (Su and Wang, 2017; Chen and Maung, 2023).

coefficient is entirely out-of-sample because only information available up to time t is used when estimating θ_t. Let the one-period ahead forecast implied by the time-varying regression be $\hat{r}_{t+1|t}$. The authors define the squared error difference (SED) between some benchmark forecast (the historical average) $\bar{r}_{t+1|t}$ and $\hat{r}_{t+1|t}$ as:

$$SED_t = (r_t - \bar{r}_{t|t-1})^2 - (r_t - \hat{r}_{t|t-1})^2. \tag{15}$$

When $SED_t > 0$, the time-varying regression forecast performs better than the benchmark. To systematically identify these periods, they further project the measure on a constant and a time trend:

$$SED_t = \gamma_{0t} + \gamma_{1t}t + v_t, \tag{16}$$

where the coefficients of the projection are again estimated with the one-sided kernel estimator. The periods where the time-varying forecast outperforms the benchmark are termed 'pockets of predictability', and can be discovered as $\widehat{SED}_t = \hat{\gamma}_{0t} + \hat{\gamma}_{1t}t > 0$. The authors use 4 familiar predictors in z_t in a univariate fashion: (i) dividend price ratio, (ii) t-bill rate, (iii) term spread, and (iv) realized variance of returns. A key finding is that the superior predictability of the time-varying predictive regressions, using the stated predictors, are local to short time intervals (pockets), and these are interspersed with longer periods of little to no predictability. The result is robust to whether we use daily or monthly daily. In other words, for most of the time, the historical average benchmark performs better than univariate time-varying predictive regressions. This is not necessarily a negative result as the authors also propose a forecast combination method that capitalizes on this fact by combining the historical average with forecasts from the time-varying regression. We return to the idea of forecast combinations later.

3.2.4. Tree Based Methods and Neural Networks - Machine Learning III

In machine learning contexts, the use of nonparametric estimation for more flexible specifications of predictive regressions raises the question on whether the linear restriction in (1) is justified in the first place. Here, we shall consider a non-linear generalization:

$$r_{t+1} = f(x_t) + u_{t+1} \tag{17}$$

where $f(\cdot)$ is an unknown and possibly non-linear function. The classical way to estimate f is by nonparametric estimation either by way of kernel or sieve regressions. There are at least two major drawbacks to these approaches: (i) if the number of predictors in x_t is large, classical nonparametric estimators will suffer from the so-called 'curse of dimensionality' and might fail to provide an estimate, and (ii) many of

these methods require the f function to be sufficiently smooth (i.e. such that their first, second or larger-order derivatives exist). Some machine learning methods require weaker conditions and may thus perform well when traditional methods fail. We consider two here: tree-based methods and neural networks.

We first discuss the basic regression tree. To simplify exposition, assume that we only have two predictors x_{1t} and x_{2t}. Next, imagine a 3-dimensional coordinate plane with (x_{1t}, x_{2t}) occupying the x and y-axes, and with r_{t+1} on the z-axis. The main idea of the regression tree is to partition the joint support of the 2-dimensional plane of (x_{1t}, x_{2t}) into regions such that the dependent variable, r_{t+1}, is constant within each region (Rossi, 2018). These regions can be obtained via a recursive binary partitioning procedure. One can then choose either x_{1t} or x_{2t} and a hypothetical splitting point. The (in-sample) predicted value of r_{t+1} on each of the two regions is called its region-mean, and the residuals from the discrepancy between the actual observation of r_{t+1} and this region-mean can be computed. Some measure of best fit can then be constructed in a very similar fashion to a least squares regression and the chosen splitting point is the one that yields the best fit. Subsequently, one or both of the regions undergo another split, yielding two more partitions per split. The process continues until some criterion is satisfied. Following the notation in Medeiros et al. (2021), we let $\hat{c}_k$ denote the sample average of r_{t+1}, within region k (i.e. region R_k). The regression tree estimator of $f(x_t)$ is given as:

$$\hat{f}(x_t) = \sum_{k=1}^{K} \hat{c}_k \hat{I}_k\{x_t\} \tag{18}$$

where $\hat{I}_k\{x_t\} = 1$, if $x_t \in R_k$, and 0 otherwise.

Boosted regression trees add on more regression trees to the original estimator (18). The first addition is trained on the in-sample residuals $(r_{t+1} - \hat{f}(x_t))$ that result from using (18), while the second addition is trained on the residuals from using the estimator with the first added tree. These additions are called boosting iterations. Rossi (2018) use a boosted regression tree with all the major predictors from Goyal and Welch (2008) in an out-of-sample forecasting and show that it outperforms many established benchmarks for forecasting stock returns (and volatility).[16]

Another variant of the regression tree is the random forest. The random forest is effectively an average of regression trees trained on bootstrap samples. Let B be the total number of bootstrap iterations. The

[16] A classical nonparametric estimator such as the kernel regression estimator in the context of (17) may not be feasible for this problem due to the relatively large number of predictors.

random forest estimator is given as:

$$\hat{f}(x_t) = \frac{1}{B}\sum_{b=1}^{B}\left(\sum_{k=1}^{K_b}\hat{c}_{k,b}\hat{I}_{k,b}\{x_t\}\right), \tag{19}$$

where objects that have the subscript b indicate estimates for a particular bootstrap iteration. In practice, since we are working with time-series data, the stationary bootstrap or block bootstrap is employed (Politis and Romano, 1994). Basak et al. (2019) use a random forest classifier to predict the direction of stock market prices.[17]

Let us now discuss neural networks. A feed-forward neural network is best described in terms of three types of layers. The input layer consists of input data or 'regressors' which correspond to our predictors. The output layer is the (in-sample) predicted value of the dependent variable. Between the input and output layer, there are hidden layers that are composed of activation units or functions that transform the input data. The layers are connected by edges (or 'arrows') that move in one direction from the input to hidden to output layers. These edges correspond to weighted aggregations. A shallow neural network is a network with only 1 hidden layer with R activation units, and can be represented as a nonparametric regression:

$$r_{t+1} = \alpha_0 + \sum_{i=1}^{R}\alpha_i g(\beta_{0,i} + x_t'\beta_i) + u_{t+1}, \tag{20}$$

where $g(\cdot)$ is an activation function (examples include the sigmoid function, the rectified linear function, and the logistic distribution function), $\beta_{0,i}$ and β_i represent the weights on the edges from the input layer (the weighted combination of predictors) that feed into the i-th activation unit, and α_i is the contribution of the i-th activation unit to the (in-sample) predicted value of $\hat{r}_{t+1}$ in the output layer. A deep neural network has $Q > 1$ hidden layers (also called hidden units). Gu et al. (2020) forecast stock-level returns using a predictor set of 94 stock-level characteristics (and interactions with 8 aggregate time-series variables) and 74 industry sector dummy variables, with regression trees and neural networks (both shallow and deep). They show that the out-of-sample performance of these methods could be double that of leading regression-based approaches. Interestingly, the authors find that shallow networks tend to deliver better performance than deep networks in their sample. Chen et al. (2024) consider a similar stock-level investigation using 178 macroeconomic time series predictors together with 46 stock-specific

[17] Tree-based classifiers are different from tree-based regressions. The main difference is that the dependent variable for a classifier is a discrete (often binary) variable. In the context of the above paper, the support of the dependent variable is $\{-1, +1\}$.

characteristics. Their approach uses a multi-step neural network (again both shallow and deep) forecasting approach, with similarly promising findings regarding the usefulness of machine learning methods.

3.2.5. Forecast Combination

We have introduced many potential models for predicting returns ranging from the simple random walk to deep neural networks. A timely question to ask at this point is *which* model is the 'best' model for out-of-sample forecasting? Furthermore, given our discussion on time-varying predictability, it is also pertinent to ask *when* they will perform well.

These questions are challenging to answer because the true underlying data generating process (DGP) of excess returns is complex and ever-changing due to factors such as institutional and policy changes, business cycle shocks, investor learning, and other structural changes (Rapach et al., 2010). It is thus highly unlikely that any single predictive model will always 'best' approximate the true DGP. This is consistent with the numerous observations above that for a given forecasting model, its predictive power is often time-varying. Hence, instead of picking just one forecast model, one could consider a combination of a suite of forecasts, ideally generated with different conditioning predictor information or from different model specifications. Such forecast combination can be viewed as a hedge against changes in the DGP that are not captured by a particular model. For a comprehensive review of forecast combination see Timmermann (2006).

If we let the one-period ahead forecast of excess returns from model i be $\hat{r}_{i,t+1|t}$, one way to combine forecasts would be the following linear weighted aggregation:

$$\hat{r}_{c,t+1|t} = \sum_{i=1}^{K} w_i \hat{r}_{i,t+1|t} \tag{21}$$

where $\hat{r}_{c,t+1|t}$ is the combined forecast of K forecasts from K different models, and where w_i is the weight attached to forecast i. The key question here is: How should we allocate the weights? The simplest way to do this is via equal weights: $w_i = 1/K$, for $i = 1, ..., K$, which corresponds to taking the average of all of the forecasts. This approach disregards the relative performance of the forecasts and runs counter to the idea that we should attach higher weights to better performing forecasts and little or no weight to inaccurate ones. Nonetheless, it is a common empirical finding that the equal weighted combination tends to dominate other more sophisticated ways of combining forecasts. This conundrum is termed the 'forecast combination puzzle', as discussed in Stock and Watson (2004). One potential reason for this is that many sophisticated combination methods require estimating a multitude of parameters, which introduces estimation noise. Another is that no one model is sophisticated enough to capture all of the

latent features of the true DGP, but by forming an average on all (or many) models, we optimally diversify away the effects of misspecification of every model. This argument is related to the idea of time varying parameters.

Given the observed superiority of the historical average forecast, many combination strategies involve combining the historical average together with other model-implied forecasts. Lin et al. (2018) run predictive regressions of corporate bond returns with 27 univariate predictors and use equal weights to combine all of their forecasts. Using their notation, we label this combined forecast $\hat{r}^{MC}_{t+1|t}$. They suggest using an additional iteration of combination with the historical average forecast, $\bar{r}_{t+1|t}$, in a regression-based combination:

$$r_{t+1} = (1-\delta)\bar{r}_{t+1|t} + \delta\hat{r}^{MC}_{t+1|t}, \tag{22}$$

where the weights of each forecast are restricted to sum to 1 and δ is estimated via OLS regression. The authors show that iterated combination with the historical average generates out-of-sample predictability that is of statistical and economic significance. The idea of using OLS regression to obtain forecast weights instead of specifying equal weights is due to Granger and Ramanathan (1984). They suggest the following specification to estimate the weights:

$$r_{t+1} = w_0 + \sum_{i=1}^{K} w_i\hat{r}_{t+1|t} + v_{t+1}, \tag{23}$$

under three different restriction schemes. The first set of restrictions is given by $w_0 = 0$ and $\sum_{i=1}^{K} w_i = 1$, the second scheme only requires $w_0 = 0$, and the final scheme has no restrictions. The last scenario is clearly the most general. Excluding an intercept term as in the first and second case can induce a biased combined forecast if the individual forecasts are biased, while this bias can be absorbed by the intercept term in the fully unrestricted case.

Farmer et al. (2023) also consider using equal weighted combinations. Recall that they estimate 4 univariate predictive regressions and check whether the forecast for each regression can beat the historical average. At time points where the model-implied forecasts outperform the historical average, they label those local periods of time as 'pockets of predictability'.[18] The authors propose the following approach to combine the forecasts from the 4 univariate regressions:

1. When the predictor is identified to be in a pocket of predictability, use the model-implied forecast, otherwise use the historical average.

[18]For each predictor, there are different sets of pockets.

2. Combine the forecasts from the 4 models using equal weights.

Hence, if none of the predictors are in any pocket, then the combined forecast is just the historical average. If one of the predictors is in a pocket, then 25% of the combined forecast is associated with the forecast implied by the corresponding predictive regression, and 75% comes from the historical average.

An intuitive way to construct combination weights would be to relate it directly to the accuracy of the forecasts. Rapach et al. (2010) consider the following construction of the weights:

$$w_{it} = \frac{\bar{\varepsilon}_{it}^{-1}}{\sum_{j=1}^{K} \bar{\varepsilon}_{jt}^{-1}}, \tag{24}$$

where the forecast error is given by:

$$\bar{\varepsilon}_{it} = \sum_{s=1}^{t-1} \theta^{t-1-s}(r_{s+1} - \hat{r}_{i,s+1|s}),$$

and $\theta < 1$ is a discount factor that places higher weight on more recent forecast errors, and lower weight on forecast errors in the distant past. The idea here is that is when forecast i exhibits good forecasting accuracy in the past, $\bar{\varepsilon}_{it}^{-1}$ is large, so a larger weight will be assigned to it. Note that the weights are continuously updated as we move forward in time (i.e. it is time-varying)[19]. When $\theta = 1$, we recover the original weights suggested by Bates and Granger (1969). They show that for $\theta = 1$, (24) is the solution to the optimization problem from minimizing $Var(\hat{r}_{c,t+1|t})$ with respect to $\{w_i\}_{i=1}^{K}$ from (21). Rapach et al. (2010) show that this combination scheme yields superior out-of-sample forecast performance relative to the historical average.

4. Methods and Models for Forecasting Volatility

The importance of volatility forecasting for investing using financial methods, and more generally in econometric modelling is obvious, as discussed above. For example, when constructing Sharpe ratios for evaluating portfolio performance, volatility forecasts are needed. Volatility is often estimated using so-called realized volatility, which is discussed in detail in the sequel. Future realized volatilities are often used in variance swaps, an important product in the volatility derivatives market. Other products that use realized volatility such as caps on variance swaps, corridor variance swaps, and options on realized volatility are also important financial instruments that are traded in financial markets. Why? Investors

[19] Regression-based forecast combinations as in (23) can also be estimated in a time-varying fashion such as with a rolling-window, or nonparametrically as in Chen and Maung (2023).

worry about future volatility risk, and hence often choose to opt for this type of contract in order to hedge against it. Realized volatility is also needed for calculation of the variance risk premium, a financial variable that has interesting implications in asset pricing. Bollerslev et al. (2009) find that the variance risk premium is able to explain time-series variation in post-1990 aggregate stock market returns with high (low) premia predicting high (low) future returns.

Jumps have a significant impact on modeling and forecasting volatility and it's realized measures. For example, when jumps are present, realized volatility is a biased estimator of integrated volatility that is defined in continuous time financial models (see below for further discussion). Thus, practitioners who are interested in modeling risks associated with continuous components of return processes, or integrated volatility, should use carefully designed realized measures of volatility that take jump effects into account.[20] Careful analyses of jumps and realized measures in the presence of jumps are crucial elements to any reasonable quantification of risk. Moreover, several authors (e.g. see Andersen et al. (2007)) have found that separation of continuous components from jump components can improve forecasts of future realized volatility. This finding should be of great interest to practitioners, especially when their objective is hedging.

Summarizing, the importance of integrated volatility (and its estimator - realized volatility), jumps and co-jumps in financial econometrics, risk management, and investing cannot be understated. However, integrated volatility is unobservable. For this reason, theorists have developed numerous relevant measures. One of the earliest is called *Realized Volatility*, as discussed in Andersen et al. (2001). However this measure does not separate jump variation from variation due to continuous components. Barndorff-Nielsen and Shephard (2004) use the product of adjacent intra-day returns to develop jump robust measures called *Bipower* and *Tripower Variations*. One of the more recent techniques for separating out the jump component is a truncation methodology which essentially eliminates returns which are above a given threshold (see Corsi et al. (2010) and Aït-Sahalia et al. (2009)). One important caveat when using high-frequency financial data in these contexts is the existence of market microstructure noise which creates a bias in the estimation procedure. Zhang et al. (2005), Zhang et al. (2006) and Kalnina and Linton (2008) solve this problem by proposing noise robust volatility estimators.

There are many methods and models used to estimate volatility by practitioners, and in the remainder of this section we discuss some of these, including simple methods like RiskMetrics, more compli-

[20] See Corradi et al. (2009) and Corradi et al. (2011) for a discussion of how to construct predictive densities for integrated volatility.

cated discrete methods including the ever popular generalized autoregressive conditional heteroskedasticity (GARCH) model, and modern methods based on parametric and nonparametric analysis of high frequency and high dimensional financial datasets.

4.1. Simple Methods and Models of Volatility

4.1.1. RiskMetrics

The RiskMetrics approach to volatility forecasting, also known as the JP Morgan method, relies on an exponentially weighted moving average model. This model implies that the forecast of today's (conditional) volatility is a weighted average of past squared returns (proxies for the unobserved true volatility)[21]:

$$\sigma^2_{t|t-1} = (1-\lambda)\sum_{i=0}^{\infty}\lambda^i z^2_{t-i-1}, \tag{25}$$

where $z_t = r_t - \bar{r}$, r_t is the return, $\bar{r}$ is the historical average of returns, and $0 < \lambda < 1$ is known as the decay or smoothing factor. We can interpret the RiskMetrics forecast as a weighted average of yesterday's actual volatility and a volatility forecast:

$$\begin{aligned}\sigma^2_{t|t-1} &= (1-\lambda)\left[z^2_{t-1} + \lambda z^2_{t-2} + \lambda^2 z^2_{t-3} + \cdots\right] \\ &= (1-\lambda)z^2_{t-1} + \lambda(1-\lambda)\left[z^2_{t-2} + \lambda z^2_{t-3} + \lambda^2 z^2_{t-4} + \cdots\right] \\ &= (1-\lambda)z^2_{t-1} + \lambda\sigma^2_{t-1|t-2}.\end{aligned} \tag{26}$$

Written in this way, it is easy to see that the exponentially weighted moving average is a restricted version of the GARCH(1,1) model to be discussed later. Specifically, (26) is equivalent to a GARCH(1,1) process with a zero intercept and coefficients that sum to 1 in the conditional volatility equation.

The decay factor λ has to be pre-specified prior to forming the forecast and the RiskMetrics approach recommends using a value of 0.94 for daily forecasts and 0.97 for monthly forecasts (see Mina et al. (2001)). An alternative approach calibrates λ by minimizing the root mean squared forecast error on a training dataset:

$$\hat{\lambda} = \text{argmin}_\lambda \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(z^2_t - \hat{\sigma}^2_{t|t-1}(\lambda)\right)^2},$$

where $\hat{\sigma}^2_{t|t-1}(\lambda)$ are different forecasts of volatility resulting from varying values of λ in (25). This optimization is easily conducted via a grid search over a predefined set of values for λ.

[21] In practice, the computation of the infinite sum in (25) is replaced by a finite summation of a specified truncation length.

Given the parsimonious forecast construction under the RiskMetrics approach, an essential question to ask is whether it can outperform more sophisticated approaches. McMillan and Kambouroudis (2009) perform a horse race to compare forecasts from GARCH-type models and the simpler RiskMetrics forecast and they report that the RiskMetrics forecasts performed well for several Asian markets while the GARCH models performed better for the G7 nations and larger Asian countries. Alexander and Leigh (1997) find a similar result. Namely, GARCH models perform more favorably compared to the exponentially weighted moving average in most of their forecasting exercises. Thus, the performance of the RiskMetrics forecast appears to be mixed when compared with more general GARCH models, but it does have the advantage of computational simplicity.

4.1.2. Autoregressive Moving Average (ARMA) Models

Suppose that we have a proxy for the unobservable return volatility such as the intraday high-low range or realized volatility (which will be discussed below) and label this proxy y_t. Our forecast of volatility can be constructed with typical time series approaches applied to this proxy. A key model here would be the following ARMA(p,q) model:

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}, \tag{27}$$

where ϵ_t is a stochastic disturbance term. This model can be estimated with either an expanding or rolling window of historical data to generate forecasts of y_{t+k}.

An early application of such models in volatility forecasting is Taylor (1986), where ARMA-type models are used to forecast future standard deviations of returns. Subsequent work applied similar models to improved measures of the volatility proxy. Hsieh (1991) forecasts daily log realized volatilities constructed from S&P500 returns in 15-minute intervals using an AR(5) model. Subsequently, Li and Hong (2011) propose a similar AR model to forecast volatility proxied by the intraday high-low range. In their forecasts of U.S. monthly realized volatility constructed from daily equity returns, French et al. (1987) use an autoregressive integrated-moving average model (ARIMA) while Schwert (1989) uses an AR(12) model, with monthly dummies.

Under the ARMA framework, it is straightforward to extend univariate volatility forecasting to the multivariate setting. The use of such multivariate systems is particularly important when the volatilities of many stocks, asset classes or markets are jointly modelled, as such systems explicitly allow for different volatilities to interact with each. Volatility spillovers between markets or stocks is one key example of such interactions. If these interactions are present in the data, univariate modeling of the volatilities will

induce an omitted variable bias which may influence forecast accuracy.

A simple model that aims to capture these effects is the vector autoregression (VAR):

$$Y_t = A_0 + \sum_{i=1}^{p} A_i Y_{t-i} + \varepsilon_t, \tag{28}$$

where Y_t is a vector of volatility proxies from different assets or markets, $A_0, ..., A_p$ are coefficient matrices, and ε_t is a stochastic disturbance term. Andersen et al. (2003) apply a VAR(5) to forecast the realized volatilities of 3 exchange rates[22], while Wang and Wan (2020) use a VAR model with structural breaks to forecast the volatilities of six highly liquid stocks.

An obstacle to the direct application of AR or VAR methods to realized volatility modeling is the issue of long memory (i.e. persistent and highly autocorrelated behavior across different time series). See Baillie (1996) and Andersen et al. (2003) for a discussion of fractionally integrated time series models that are relevant in this context. Interestingly, it is possible to approximate a stationary fractionally integrated time series with an AR model with increasing lag order. Poskitt (2007) provides several theoretical conditions to guarantee the validity of such an approximation while Wang et al. (2013) extend the results to fractionally integrated processes with structural breaks. For an alternative approach, Bauwens et al. (2023) study the conditions under which a VAR(1) can sufficiently account for long memory in the individual variables and propose a restricted estimation approach (via penalization) of the VAR(1) such that the conditions are satisfied. Another model that is relevant in this context is the heterogeneous autoregressive model of Corsi (2009).

4.2. Advanced Methods and Models of Volatility

4.2.1. Discrete Time ARCH and GARCH Models

At the time of the development of GARCH models to forecast volatility in the 1980s and 1990s there were a number of stylized facts that were front and center in the minds of practitioners and researchers. Some of these included the following:

Leptorkurtosis: Asset returns were noted by Mandelbrot (1963) and Fama (1965) to have fat tails, indicating the importance of using non-normal distributions to model their dynamics. Fama (1965) shows evidence of excess kurtosis in the distribution of stock returns. Engle and Gonzalez-Rivera (1991) introduce a semi-parametric volatility model, which allows for generic return distributions.

[22]The time series used have been fractionally differenced to deal with the problem of long memory.

Volatility Clustering and Persistence: By observing cotton prices, Mandelbrot (1963) stressed that ".... large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes...". The persistence of shocks to the conditional variance of stock returns seems to be clear. The interpretation of this persistence as well as how long the shocks persist is crucial in specifying the "correct" dynamics. Poterba and Summers (1986) note that volatility shocks may affect the entire term structure, associated risk premia, and investment in long-lived capital goods.

Broadly speaking, volatility persistence is an important feature that pertains to models with time varying and codependent variance structures. Black and Scholes (1973) write that "...there is evidence of non-stationary in the variance. More work must be done to predict variances using the information available.". Since their paper, numerous autoregressive conditional heteroskedasticity, volatility and stochastic volatility models have been developed.

Leverage Effects: Black (1976) observed that changes in stock prices seem to be negatively correlated with changes in stock volatility. Volatility seems to increase after bad news and decrease after good news. Schwert (1989) and Schwert (1990) presents empirical evidence that stock volatility is higher during recessions and financial crises. Christie (1982) discusses economic mechanisms that explain this effect. Specifically, reductions in equity value raise the riskiness of firms, as implied by debt to equity ratios, and therefore lead to increases in future volatility. For modeling, Nelson (1991) suggests a new model that captures the asymmetric relation between returns and changes in volatility.

Co-movement in Volatilities: This feature was also first commented on by Black (1976). He points out the commonality in volatility changes across stocks. When stock volatilities change, they all tend to change in the same direction. This suggests that (few) common (unobserved or missing) factors might be specified when modelling individual asset return volatility.

ARCH and later GARCH models were specified to adhere as closely as possibly to the above stylized facts. Consider the autoregressive conditional heteroskedasticity (ARCH) model of Engle (1982) and the Generalized ARCH (GARCH) model of Bollerslev (1986), as well as related models. These models are very well known, but we briefly discuss them in order to trace the evolution to modern continuous-time stochastic volatility models.

Let X_t be a financial asset return, say, and F_{t-1} denotes a filtration of all information through time $t-1$. The prototypical autoregressive conditional heteroskedasticity (ARCH) model has:

$$X_t = \varepsilon_t \sigma_t$$

where ε_t is a stochastic disturbance term. Here,

$$\varepsilon_t \sim i.i.d \text{ with}$$

$$E(\varepsilon_t) = 0 \text{ and } Var(\varepsilon_t) = 1 \text{ and}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2.$$

where the α's are constants to be estimated. In the more general case:

$$X_t | F_{t-1} \sim N(Z_t\beta, \sigma_t^2),$$

$$\sigma_t^2 = h(\varepsilon_{t-1}, \varepsilon_{t-2,\ldots,}\varepsilon_{t-p}, \alpha), \text{ and}$$

$$\varepsilon_t = X_t - Z_t\beta,$$

where Z_t may contain lags of X_t. and F_{t-1} is a data filtration, as discussed above. If the function h contains current and lagged X's, then

$$\sigma_t^2 = h(\varepsilon_{t-1}, \varepsilon_{t-2,\ldots,}\varepsilon_{t-p}, x_t, x_{t-1,\ldots,}x_{t-p}, \alpha).$$

In this class of models, ARCH(p) is the most popular one, and has the following specification:

$$X_t | F_{t-1} \sim N(Z_t\beta, \sigma_t^2),$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \ldots \alpha_p \varepsilon_{t-p}^2.$$

Engle (1982) proposes a convenient estimation and testing methodology for this model using maximum likelihood. He shows that α and β can be estimated separately under some regularity conditions.[23] To capture the trade-off between risk and expected return, Engle et al. (1987) introduce ARCH in mean, or ARCH-M models, where: Let

$$X_t = g(Z_{t-1}, \sigma_t^2; b) + \varepsilon_t,$$

with $g(\cdot)$ being some appropriately defined function. The appealing feature of this model is that the conditional mean, μ_t, is a function of the variance, i.e. $\mu_t = g(Z_{t-1}, \sigma_t^2; b)$. This helps us to directly model the risk-return relationship, and has important implications for predicting the conditional mean function, since the conditional volatility enters therein. In practice, many papers set $g(\cdot)$ to be a linear or logarithmic function.

[23] For details, see Sections 4 and 5 in Engle (1982).

An important improvement to these models is made by Bollerslev (1986), where the ARCH model is generalized to the Generalized ARCH (GARCH) model. As noted in Bollerslev (1986), the extension from ARCH to GARCH is similar to the extension in time series modelling of an AR to an ARMA model. Specifically, as in the case of the ARCH model, let ε_t be the innovation in a linear regression

$$\varepsilon_t = X_t - Z_t'\beta,$$

where β is a vector of parameters. Then the GARCH (p,q) specification is given by

$$\begin{aligned}
\varepsilon_t | F_{t-1} &\sim N(0, \sigma_t^2), \\
\sigma_t^2 &= \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2, \\
\varepsilon_t &= X_t - Z_t\beta,
\end{aligned}$$

where p and q denote lag orders, and

$$\begin{aligned}
&p \geq 0, q > 0, \\
&\alpha_0 > 0,\ \alpha_i \geq 0,\ i = 1, ..., q,\ \text{and} \\
&\beta_i \geq 0,\ i = 1, ..., p.
\end{aligned}$$

The difference between the GARCH and ARCH models is that the former includes lagged conditional variances, allowing for a potentially better 'fit' of the model, and achieving an impressive level of parsimony, given that the GARCH(1,1) specification is often found to perform the best in empirical applications. Bollerslev (1986) discusses maximum likelihood estimation as well as testing procedures for GARCH (p,q) models. The most successful model, empirically, is the GARCH(1,1) model. An important variant of this model, the so-called Integrated GARCH or IGARCH model is discussed in Engle and Bollerslev (1986). Under IGARCH, $\sum_{i=1}^{q} \alpha_i + \sum_{i=1}^{p} \beta_i = 1$, which implies a unit root in the volatility equation.

In other key papers, Nelson (1990) and Nelson (1991) discuss the use of EARCH (i.e., exponential ARCH) to approximate continuous time processes. Nelson (1991) points out that the GARCH model has several limitations in empirical applications to financial markets. For instance, in the GARCH model, volatility responds symmetrically to positive and negative residuals and therefore does not explain the stylized leverage effect. In lieu of this, Nelson (1991) proposes the EARCH model, which is specified as follows:

$$X_t = \sigma_t \varepsilon_t, \text{ and } \varepsilon_t \sim i.i.d \text{ with}$$

$$E(\varepsilon_t) = 0 \ , \ Var(\varepsilon_t) = 1,$$

and

$$\ln(\sigma_t^2) = \alpha_t + \sum_{k=1}^{\infty} \beta_k g(\varepsilon_{t-k}), \qquad \beta_1 \equiv 1$$

where $\{\alpha_t\}_{t=-\infty,\infty}$ and $\{\beta_k\}_{k=1,\infty}$ are parameters. The choice for the functional form of $g(\cdot)$ is $g(\varepsilon_t) = \theta\varepsilon_t + \gamma(|\varepsilon_t| - E|\varepsilon_t|)$. This set-up allows the conditional variance process to respond asymmetrically to rises and falls in stock prices. It is straightforward to verify this as when ε_t is positive $g(\varepsilon_t) = (\theta+\gamma)\varepsilon_t - \gamma E(|\varepsilon_t|)$ and when ε_t is negative $g(\varepsilon_t) = (\theta - \gamma)\varepsilon_t - \gamma E(|\varepsilon_t|)$. In each case, the $g(\varepsilon_t)$ is a linear function with a different slope. In addition, Nelson (1991) points out that while for GARCH it is difficult to verify the persistence of the shocks to the variance, in the EARCH model the stationarity and ergodicity of the logarithm of the variance process is easily checked. Other modifications of the GARCH (1,1) model include the GJR model proposed by Glosten et al. (1993). This model imposes structure that induces asymmetry in shocks to returns in a different way. Namely, they define:

$$\sigma_t^2 = \omega + \alpha\varepsilon_t^2 + \gamma\varepsilon_t^2 1_{\{\varepsilon_t \geq 0\}} + \beta\sigma_{t-1}^2$$

Note that when $\gamma < 0$, positive return shocks increase volatility less than negative shocks. For a complete list and discussion of these and a whole host of related models, see Bollerslev (2008), where he provides a Glossary to ARCH. For models with multivariate specifications (see Bollerslev et al. (1988)).

An interesting aspect of the volatility literature is the connection between discrete time and continuous time models. In the case of constant volatility, the classical result by Cox and Ross (1976) shows that the limiting form of the jump process

$$dX_t = \mu X_t dt + cX_t dN_t(\lambda)$$

as $\lambda \to 0$ is the diffusion process

$$dX_t = \mu X_t dt + \sigma X_t dW_t$$

where σ is a function of c. $N_t(\lambda)$ is a continuous time Poisson process with intensity λ, (i.e., dN_t is the number of jumps of X_t during dt and is Poisson-distributed with parameter λdt), cX_t is the jump amplitude, and W_t is a standard Brownian motion. Using this setup, Nelson (1990) bridges the gap between discrete and continuous time stochastic volatility models by using AR(1) Exponential ARCH and GARCH (1,1) models as approximations for continuous time processes of the variety discussed in the next two sections.

4.2.2. Nonparametric Realized Measures of Integrated Volatility

The above discussion serves as a natural starting point for our discussion of continuous time methods used for volatility forecasting. As mentioned above, these methods often involve utilizing big data (e.g., high frequency and high dimensional financial data).

Using notation very similar to that used in earlier sections, we begin by characterizing the log-price of a financial asset at continuous time t, as Y_t. It is assumed that the log-price is a Brownian semi-martingale process with jumps and that can be denoted as follows:[24]:

$$Y_t = Y_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s + J_t. \tag{29}$$

In equation (29), μ_s the a predictable drift process, the diffusion term σ_s is a cádlág process, W_s is a standard Brownian motion and J_t is a pure jump process. J_t can be defined as the sum of all discontinuous log price movements up to time t,

$$J_t = \sum_{s \leq t} \Delta Y_s.$$

When this jump component is a finite activity compound Poisson jump process, then:

$$J_t = \sum_{j=1}^{N_t} \xi_j,$$

where N_t is a Poisson process with intensity λ, the jumps occur at the corresponding times given as $(\tau_j)_{j=1,..,N_t}$ and ξ_j is an $i.i.d$ random variable measuring the size of jumps at time τ_j. The finite activity jump assumption has been widely used in the financial econometrics literature, for example. Y_t can be decomposed into a continuous Brownian component, Y_t^c, and a discontinuous (jump) component, Y_t^d. The 'true variance' of Y_t can be given as,

$$QV_t = [Y,Y]_t = [Y,Y]_t^c + [Y,Y]_t^d,$$

where QV stands for quadratic variation. The variation due to the continuous component is:

$$[Y,Y]_t^c = \int_0^t \sigma_s^2 ds,$$

and the variation due to the discontinuous jump component is

$$[Y,Y]_t^d = \sum_{j=1}^{N_t} \xi_j^2.$$

[24]We follow the setup and notation used in Corradi et al. (2011)

Integrated volatility, which is the continuous part of QV, is denoted as:

$$IV_t = \int_{t-1}^{t} \sigma_s^2 ds, \quad t = 1, ..., T,$$

where IV is the integrated volatility at day t. Since IV is unobservable, different realized measures (estimators) of integrated volatility are used in empirical applications when forecasting volatility.

A wrinkle to this setup is the presence of market frictions in high frequency financial data, as has been documented in recent literature. To take care of this, the observed log price process, say X_t can be given as

$$X_t = Y_t + \epsilon_t$$

where Y_t is the latent log price and ϵ_t captures market microstructure noise.

Now, consider M equi-spaced intradaily observations for each of T days for X_t, which yields a total of MT observations. Namely, define:

$$X_{t+j/M} = Y_{t+j/M} + \epsilon_{t+j/M}, \quad t = 0, .., T, \quad j = 1, .., M, \tag{30}$$

where $\epsilon_{t+j/M}$ follows a zero mean independent process. Finally, the intraday return or increment of process X_t is defined as:

$$\Delta_j X = X_{t+(j+1)/M} - X_{t+j/M}$$

The noise containing realized measure, $RM_{t,M}$ of the integrated volatility is computed using process X_t given in (30) and can be expressed as the sum of IV and measurement error N, i.e.

$$RM_{t,M} = IV_t + N_{t,M}.$$

RM can be used to estimate IV if the k^{th} moment of the measurement error decays to zero at a fast enough rate or there exists a sequence b_M with $b_M \to \infty$ such that $E(|N_{t,M}|^k) = O(b_M^{-K/2})$, for some k $\geq$2.

Using this setup, and noting that parametric models have been shown to be misspecified when used to capture volatilities implied by option pricing and other financial return variables, it is perhaps not surprising that substantial effort has been made to construct model free estimators of volatility. This is feasible, given availability of high frequency data, as first explained in Andersen et al. (2001). The measure introduced in this paper, termed *Realized Volatility* (RV), is constructed by summing over intra-day squared returns. The authors show that RV is an error free estimator of integrated volatility (i.e., IV can be consistently estimated using RV) in the absence of noise and jumps. On the other hand,

when the sampling frequency of the data is relatively high, microstructure noise creates a bias in the volatility estimation procedure. Zhang et al. (2005), Zhang et al. (2006) and Kalnina and Linton (2008) solve this problem with microstructure noise robust estimators based on sub-sampling with multiple time scales. Barndorff-Nielsen et al. (2008) and Barndorff-Nielsen et al. (2011) on the other hand, use kernel based estimators to account for the microstructure noise. When estimating integrated volatility in the presence of jumps, the jump components should be separated from quadratic variation. Barndorff-Nielsen and Shephard (2003) and Barndorff-Nielsen and Shephard (2004) provide asymptotically unbiased integrated volatility estimators, called bipower and tripower variation, which are robust to the presence of jumps. Aït-Sahalia et al. (2009) propose a threshold method to identify and truncate jumps and develop another consistent nonparametric jump robust estimator of integrated volatility. Corsi et al. (2010) introduces threshold bipower variation by combining the concepts from Barndorff-Nielsen and Shephard (2003) and Mancini (2009). Jacod et al. (2014) estimate local volatility by using the empirical characteristic function of returns to remove bias due to jump variation. When combining both jumps and microstructure noise in the price process, Fan and Wang (2007) propose a wavelet-based multi-scale approach to estimating integrated volatility. Podolskij et al. (2009) design modulated bipower variation, an estimator that filters the impact of microstructure noise and then use bipower variation for volatility estimation. Andersen et al. (2012) use the concept of 'nearest neighbor truncation' to establish jump and noise robust volatility estimators. Brownlees et al. (2016) create truncated two scaled realized volatility by adopting a jump signaling indicator as in Mancini (2009) and noise robust sub-sampling as in Zhang et al. (2005). In addition to the above mentioned work, discussion regarding nonparametric estimation of integrated volatility and functionals of volatility can also be found in Barndorff-Nielsen et al. (2006), Mykland and Zhang (2009), Todorov and Tauchen (2012), Hautsch and Podolskij (2013), Jacod et al. (2013), Jing et al. (2014) and Jacod et al. (2019). In the next section, we review some of the most commonly used realized measures.[25]

Realized Volatility (RV) as developed in Andersen et al. (2001) is one of the first empirical measures that used high-frequency intra-day returns to compute daily return variability without having to explicitly model the intra-day data. The authors show that under suitable conditions RV is an unbiased and highly efficient estimator of QV. By extension it can be shown that in the absence of jumps or when jumps populate the data infrequently, RV converges in probability to IV as $M \longrightarrow \infty$.

[25]For further details, refer to Mukherjee et al. (2020).

Before defining RV, it should be noted that RV is widely used in HAR-RV forecasting models, such as those first developed by Corsi (2009). The basic HAR-RV model is specified as follows:

$$\phi(RV_{t,t+h}) = \beta_0 + \beta_d\phi(RV_t) + \beta_w\phi(RV_{t-5,t}) + \beta_m\phi(RV_{t-22,t}) + \epsilon_{t+h},$$

where daily RV, $t+h$ days ahead, is forecast using daily ($\phi(RV_t)$), weekly ($\phi(RV_{t-5,t})$), monthly ($\phi(RV_{t-22,t})$) RV measures, and ϵ_{t+h} is a stochastic disturbance term. Here, $\phi(\cdot)$ is usually set equal to a linear, square root, or log function, so that in its simplest form this is just a linear regression of future RV on past daily, weekly and monthly RV. Volatility forecasting models using this sort of regression can also be constructed by replacing RV with other realized measures such as those discussed below.

So how do we estimate RV? This is the simplest realized measure, and is constructed by summing squared intraday returns, as mentioned above. Namely, define:

$$RV_{t,M} = \sum_{j=1}^{M-1} (X_{t+(j+1)/M} - X_{t+j/M})^2.$$

Realized Bipower Variation (BPV) is an estimator originally introduced in Barndorff-Nielsen and Shephard (2004). In their paper, these authors demonstrate how to untangle the continuous component of quadratic variation from its discontinuous component (jump component), using their BPV estimator, which was one of the first asymptotically unbiased estimators of IV which was robust to the presence of price jumps. The BPV realized measure takes the following form:

$$BPV_{t,M} = (\mu_1)^{-2} \sum_{j=2}^{M-1} |\Delta_j X||\Delta_{j-1} X|,$$

where $\Delta_j X$ is the same as in (4.2.2) and $\mu_1 = 2^{\frac{1}{2}} \frac{\Gamma(1)}{\Gamma(\frac{1}{2})}$.

Tripower Variation (TPV) is another consistent estimator of IV in the presence of finite activity jumps. However since BPV is subject to finite sample jump distortions or upward bias, BPV was proposed by Barndorff-Nielsen and Shephard (2004). This realized measure utilizes products of the (lower order) power of three adjacent intra-day returns, and is theoretically more efficient than BPV, also it is also more vulnerable to the effects of microstructure noise than BPV. TPV is defined as follows:

$$TPV_{t,M} = (\mu_{\frac{2}{3}})^{-3} \sum_{j=3}^{M-1} |\Delta_j X|^{2/3} |\Delta_{j-1} X|^{2/3} |\Delta_{j-2} X|^{2/3},$$

where $\Delta_j X$ is defined above and $\mu_{\frac{2}{3}} = 2^{\frac{1}{3}} \frac{\Gamma(\frac{5}{6})}{\Gamma(\frac{1}{2})}$.

Two Scale Realized Volatility (TSRV) was developed after it was found that when the sampling interval of the asset prices is small, microstructure noise issues become more prominent and *Realized Volatility*

ceases to function as a robust volatility estimator. Due to the bias introduced by the market microstructure noise in the finely sampled data, longer time horizons were initially preferred by practitioners when contructing realized measures. Why is this? Because it was found that ignoring microstructure noise works well for intervals longer than 10 minutes. However sampling over lower frequencies does not fully correct for the effects of noise on volatility estimation. As a solution, TSRV was introduced by Zhang et al. (2005). This estimator combines estimators obtained over two time scales, avg and M, and is an unbiased and consistent, microstructure noise robust estimator of IV in the absence of jumps. It takes the following form:

$$TSRV_{t,M} = [X,X]^{avg} - \frac{1}{K}[X,X]^{M},$$

where

$$[X,X]^{m_i} = \sum_{j=1}^{m_i-1} (X_{t+((j+1)K+i)/M} - X_{t+(jK+i)/M})^2, \;\; i = 1,..,K \;\; \& \;\; m_i = \frac{M}{K},$$

$$[X,X]^{avg} = \frac{1}{K}\Sigma_{i=1}^{K}[X,X]^{m_i},$$

and

$$[X,X]^{M} = \sum_{j=1}^{M-1} (X_{t+(j+1)/M} - X_{t+j/M})^2,$$

where $K = cM^{2/3}$ is the number of subsamples, $\frac{M}{K}$ is subsample size, $c > 0$ is a constant and M is the number of equi-spaced intraday observations.

Multi-Scale Realized Volatility (MSRV) was developed because although the TSRV estimator has many desirable properties, is not efficient. The rate of convergence for TSRV to the true volatility (IV) in the absence of jumps is of the order $M^{-1/6}$. The MSRV estimator proposed in Zhang et al. (2006) is a microstructure noise robust measure which converged to IV (in the absence of jumps) at the rate of $M^{-1/4}$. While $TSRV$ uses two time scales, $MSRV$ uses N different time scales. $MSRV$ is defined as follows:

$$MSRV_{t,M} = \sum_{n=1}^{N} a_n [X,X]^{(M,K_n)}, \;\; n = 1,..,N,$$

where

$$a_n = 12\frac{n}{N^2}\frac{n/N - 1/2 - 1/(2N)}{1 - 1/N^2}, \;\; \sum_{n=1}^{N} a_n = 1 \;\;\; , \;\; \sum_{n=1}^{N} a_n/n = 0,$$

and

$$[X,X]^{(M,K_n)} = \frac{1}{K_n}\sum_{l=1}^{K_n}\sum_{j=1}^{m_{n,l}-1} (X_{t+((j+1)K_n+l)/M} - X_{t+(jK_n+l)/M})^2.$$

Here $l = 1,..,K_n$ and $m_{n,l} = \frac{M}{K_n}$. We take $N = 3, K_1 = 1, K_2 = 2, K_3 = 3$.

Realized Kernel (RK) volatility is an estimator introduced in Barndorff-Nielsen et al. (2008). As the name suggests, it is a realized kernel type consistent measure of IV in the absence of jumps. It is robust to endogenous microstructure noise and for particular choices of weight functions it can be asymptotically equivalent to $TSRV$ and $MSRV$ estimators, or even more efficient. RK is constructed using the following equation:

$$RK_{t,M} = \gamma_0(X) + \sum_{h=1}^{H} \kappa(\frac{h-1}{H})\{\gamma_h(X) + \gamma_{-h}(X)\},$$

where

$$\gamma_h(X) = \sum_{j=1}^{M-1} (X_{t+(j+1)/M} - X_{t+j/M})(X_{t+(j+1-h)/M} - X_{t+(j-h)/M}).$$

Here c is a constant. One can use the Turkey-Hanning$_2$ kernel with $\kappa(x) = sin^2\{\pi/2(1-x)^2\}$ and $H = cM^{1/2}$.

Truncated Realized Volatility (TRV) is one of the first volatility measures that tries to estimate IV by identifying when price jumps greater than an adequately defined threshold occur, as discussed in Aït-Sahalia et al. (2009). The truncation level for the jumps is chosen in a data-driven manner. The price jump robust measure is:

$$TRV_{t,M} = \sum_{j=1}^{M-1} |\Delta_j X|^2 1_{\{|\Delta_j X| \leq \alpha \Delta_M^{\varpi}\}},$$

where

$$\alpha = 5\sqrt{\sum_{j=1}^{M-1} |\Delta_j X|^2 1_{\{|\Delta_j X| \leq \Delta_M^{1/2}\}}}.$$

Here one might set $\varpi = 0.47$, for example, and $\Delta_M = 1/M$.

Modulated Bipower Variation (MBV) is introduced in Podolskij et al. (2009) and consistently estimates IV, is robust to market microstructure noise, and is also robust to finite activity jumps. This realized measure is constructed as follows:

$$MBV_{t,M} = \frac{(c_1 c_2/\mu_1^2) mbv_{t,M} - \vartheta_2 \hat{\omega}^2}{\vartheta_1},$$

where

$$\vartheta_1 = \frac{c_1(3c_2 - 4 + max((2-c_2)^3, 0))}{3(c_2-1)^2}, \quad \vartheta_2 = \frac{2min((c_2-1), 1)}{c_1(c_2-1)^2},$$

$$mbv_{t,M} = \Sigma_{b=1}^{B} |\bar{X}_b^{(R)}||\bar{X}_{b+1}^{(R)}|$$

and

$$\bar{X}_b^{(R)} = \frac{1}{M/B - R + 1} \Sigma_{j=(b-1)M/B}^{bM/B-R} (X_{t+(j+R)/M} - X_{t+j/M}).$$

Here $c_1 = 2$, $c_2 = 2.3$, $R \approx c_1 M^{0.5}$, $B = 6$, $\mu_1 = 0.7979$, and $\hat{\omega}^2 = \frac{1}{2M} RV_{t,M}$.

Subsampled Realized Kernel (SRK) is an estimator due to Barndorff-Nielsen et al. (2011) and is constructed by combining the concepts of subsampling (Zhang et al. (2005)) and realized kernels (Barndorff-Nielsen et al. (2008)). The main benefit of subsampling in our context is that it can reduce inefficiency stemming from the poor selection of kernel weights that might be the case when using realized kernel estimators. SRK is constructing as follows:

$$SRK_{t,M} = \frac{1}{S} \sum_{s=1}^{S} K^s(X),$$

where

$$K^s(X) = \gamma_0^s(X) + \sum_{h=1}^{H} \kappa(\frac{h-1}{H}) \{\gamma_h^s(X) + \gamma_{-h}^s(X)\},$$

$$\gamma_h^s(X) = \sum_{j=1}^{\frac{M}{S}} x_j^s x_{j-h}^s,$$

and

$$x_j^s = X_{t+(j+\frac{s-1}{S})/M} - X_{t+(j+\frac{s-1}{S}-1)/M},$$

Here one can use the smooth Turkey-Hanning$_2$ kernel function with $\kappa(x) = sin^2\{\pi/2(1-x)^2\}$, $S = 13$, and $H = 3$.

MedRV and MinRV are estimators that are alternatives to *Realized Bipower Variation* and *Tripower Variation*, as discussed in Andersen et al. (2012). Both are robust to jumps and/or microstructure noise and use 'nearest neighbor truncation'. The basic concept behind these measures is that neighboring returns control the level of truncation of absolute returns. $MinRV$ compares and takes the minimum of two adjacent absolute returns, $MedRV$ takes the median of three adjacent absolute returns and carries out two-sided truncation. Unlike the typical truncated realized measures as in Corsi et al. (2010), these new measures do not require the selection of an ex-ante threshold. They are defined as follows:

$$MinRV_{t,M} = \frac{\pi}{\pi - 2} (\frac{M}{M-1}) \sum_{j=1}^{M-1} min(|\Delta_j X|, |\Delta_{j+1} X|)^2$$

and

$$MedRV_{t,M} = \frac{\pi}{6 - 4\sqrt{3} + \pi} (\frac{M}{M-2}) \sum_{j=2}^{M-1} med(|\Delta_{j-1} X|, |\Delta_j X|, |\Delta_{j+1} X|)^2,$$

where $\Delta_j X$ is defined above.

All of the above estimators are nonparametric, in the sense that parametric models need not be specified, although when constructing forecasts, one still specifies a parametric forecasting model, as discussed above in the context of HAR-RV forecasting regressions.

In general, there are certain advantages to specifying and estimating parametric models at all stages when forecasting and simulating volatility. One obvious advantage is that pathwise simulation and forecasting that retains specific dependence is feasible only under parametric model specification. For this reason, we conclude this section by discussing widely used stochastic volatility models.

4.2.3. Parametric Continuous Time Models

Many high-frequency financial time series appear to be a mixture of sudden relatively large changes and smooth small changes. This image suggests modeling a high-frequency financial time series as a mixture of a discrete jump process and a continuous diffusion process. A jump-process J_t is a discrete process specified by a distribution, ν, for the magnitudes of the jumps and a distribution, $\lambda(X_t)$, for the intensity with which jumps occur, as discussed above. A jump-diffusion process is the sum of a continuous diffusion process and a jump process,

$$dX_t = \mu(X_t, t, \theta)dt + \sigma(X_t, t, \theta)dW + dJ_t$$

One pioneering work which incorporates jumps into continuous time processes is Merton (1976), where he models the continuous component of the log price process to be Gaussian as in the case of geometric Brownian motion. The magnitude of jumps also follows a Gaussian distribution, and jumps follow Poison distribution in his paper. Newer developments in this area do not "append" a "discrete" jump process onto the diffusion, but instead specify the jumps using other means, such as via the use of Levy processes.

A natural refinement of this model is the stochastic volatility model. These models were first introduced by Harvey et al. (1994) in discrete time. Stochastic volatility implies that unobserved volatility follows another stochastic process. For example, one specification could be

$$dX_t = (\alpha + \beta X_t)dt + \sigma dW_{1t},$$

where the volatility process follows:

$$d\sigma_t^2 = \kappa(\vartheta - \sigma_t^2)dt + \delta\sigma_t dW_{2t},$$

with $Cov(dW_{1t}, dW_{2t}) = \rho dt$.

CIR type Stochastic Volatility Model: Andersen and Lund (1997) estimate the generalized Cox-Ingersoll-Ross model discussed above with:

$$dX_t = \kappa_1(\alpha - X_t)dt + \sigma_t X_t^{\beta} dW_{1t},$$

$$d\log\sigma_t^2 = \kappa_1(\alpha - \log\sigma_t^2)dt + \delta dW_{2t},$$

where the $W(\cdot)$ terms are Brownian motions, all all other terms are constants except for the price and voltility measures, X_t and σ_t.

Mixed Stochastic Volatility, Jump Diffusion Model: These models arise from the application of spectral GMM in Chacko and Viceira (2003), where the return/volatility process is specified as:

$$dX_t = (\mu - \frac{\sigma_t^2}{2})dt + \sigma_t dW_{1t} + [\exp(J_u) - 1]dN_u(\lambda_u) + [\exp(-J_d) - 1]dN_d(\lambda_d)$$
$$d\sigma_t^2 = \kappa(\alpha - \sigma_t^2)dt + \delta\sigma_t dW_{2t},$$

where λ_u, λ_d are jump intensity parameters and are constant, and where J_u and J_d >0 are stochastic jump magnitudes that follow exponential distributions, i.e.

$$f(J_u) = \frac{1}{\eta_u}\exp(\frac{-J_u}{\eta_u}),$$
$$f(J_u) = \frac{1}{\eta_d}\exp(\frac{-J_d}{\eta_d}).$$

In the option pricing literature, many models are nested in the following data generating process which, allows for jumps in both equations

$$dX_t = \mu_t dt + \sigma_t X_t dW_{1t} + dJ_{1t}$$
$$d\sigma_t^2 = \kappa(\alpha - \sigma_t^2)dt + \delta\sigma_t(\rho dW_{1t} + \sqrt{1-\rho^2}dW_{2t}) + dJ_{2t}$$

where W_{1t} and W_{2t} are independent Brownian motions, and J_{1t} and J_{2t} are jump processes. Popular models that are nested in this class include those in Heston (1993), Bates (2000), Chernov and Ghysels (2000). and Pan (2002).

A natural way to forecast volatility in the above parametric models involves constructing predictive volatility densities using simulated data, after estimating the parameters of the models. This is discussed in detail in Corradi et al. (2009) and Corradi et al. (2011).

4.2.4. Big Data Methods in Continuous Time, Factor Estimation and Volatility Prediction Using The LASSO and the Elastic Net for Variable Selection - Machine Learning IV

This section summarizes the results reported in Cheng et al. (2021) on volatility forecasting using factors estimated with high frequency and high dimensional financial datasets. The basic idea is to use realized measures (RMs) in a continuous time factor representation from which factors are extracted and subsequently incorporated in factor augmented forecasting regressions. While the setup used is similar in

some regards to the setup used in the previous sections, key underlying features are different, as is the notation used. For this reason, we start at the very beginning by considering a d-dimensional process, X, consisting of d asset log-prices. Assume that X follows an Itô-semimartingale defined on the filtered probability space $\left(\Omega, \mathbb{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P}\right)$, and has the following representation:

$$\begin{aligned} X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s \\ + \int_0^t \int_{\{|x|\leq \epsilon\}} x(\mu - \nu)(ds, dx) + \int_0^t \int_{\{|x|\geq \epsilon\}} x\mu(ds, dx), \end{aligned}$$

where b_t is the instantaneous drift term, σ_t is the spot covariance, and both are adapted, càdlàg, and locally bounded. Additionally, W_t is a multidimensional standard Brownian motion, μ is a random jump measure with compensator ν, and $\epsilon > 0$ is an arbitrary threshold. For more details on Itô-semimartingales and continuous-time asset price modeling, see Aït-Sahalia and Jacod (2014) and the references cited therein.

As discussed above, various realized measures have been developed to estimate latent volatility on a fixed interval $[0, T]$, using high-frequency intraday data. For instance, recall that realized volatility, one of the most widely known measures, is given by:

$$\text{RV}_t = \sum_{i=1}^{\lfloor t/\Delta_n \rfloor} \left(\Delta_i^n X^j\right)^2, \quad \forall t \in [0, T], \quad j = 1, ..., d, \tag{31}$$

where $\lfloor \cdot \rfloor$ is the floor function and $\Delta_i^n X^j = X^j_{i\Delta_n} - X^j_{(i-1)\Delta_n}$ is the i^{th} intraday return for j^{th} asset in X, with Δ_n defined as an equally-spaced sampling interval that shrinks to zero. It is well-known that when asset prices are continuous on a fixed interval, $[0, T]$, we have that:

$$\sum_{i=1}^{\lfloor t/\Delta_n \rfloor} \left(\Delta_i^n X^j\right)^2 \xrightarrow{\mathbb{P}} \int_0^t \left(\sigma_s^j\right)^2 ds, \quad \forall t \in [0, T], \quad j = 1, ..., d,$$

as $\Delta_n \to 0$, where σ_s^j is the spot volatility for j^{th} asset.

However, when asset prices are discontinuous on $[0, T]$:

$$\sum_{i=1}^{\lfloor t/\Delta_n \rfloor} \left(\Delta_i^n X^j\right)^2 \xrightarrow{\mathbb{P}} \int_0^t \left(\sigma_s^j\right)^2 ds + \sum_{0\leq s\leq t} \left(\Delta X_s^j\right)^2, \quad \forall t \in [0, T], \quad j = 1, ..., d,$$

where $\Delta X_s^j = X_s^j - X_{s-}^j \neq 0$, if and only if the j^{th} asset, X^j, jumps at time s. To separate integrated volatility from jump variation, one might use the threshold technique developed by Mancini (2001, 2009) to construct truncated realized volatility (TRV), defined as:

$$\text{TRV}_t = \sum_{i=1}^{\lfloor t/\Delta_n \rfloor} \left(\Delta_i^n X^j\right)^2 \mathbf{1}_{\{|\Delta_i^n X^j| \leq c\Delta_n^{\varpi}\}} \xrightarrow{\mathbb{P}} \int_0^t \left(\sigma_s^j\right)^2 ds, \tag{32}$$

for some $\varpi \in (0, 1/2)$, or use the multipower variation (MPV) estimator developed by Barndorff-Nielsen and Shephard (2004), where:

$$\text{MPV}_t = \Delta_n^{1-p^+/2} \sum_{i=1}^{\lfloor t/\Delta_n \rfloor - k+1} |\Delta_i^n X^j|^{p_1} \cdots |\Delta_{i+k-1}^n X^j|^{p_k} \xrightarrow{\mathbb{P}} m_{p_1} \cdots m_{p_k} \int_0^t |\sigma_s^j|^{p^+} ds, \tag{33}$$

with $p_1, p_2, \ldots, p_k \geq 0$, $p^+ = p_1 + \cdots + p_k$ and $m_p = \mathbb{E}[|\mathcal{N}(0,1)|^p]$. One can also combine these two methods and use a truncated multipower variation estimator (see Corsi et al. (2010)). This allows for different components of quadratic variation to be separately analyzed.

To facilitate the analysis of large dimensional datasets, we further assume that the continuous part of asset log-prices panel follows a continuous-time factor model on $[0, T]$. Namely, define $Y_t := X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s$ as the continuous part of X, and assume the following factor structure for Y_t:

$$Y_t = \Lambda_t F_t + Z_t, \tag{34}$$

where F_t is an r-dimensional ($r < d$) unobservable common factor, Z_t is an idiosyncratic component, and Λ_t is a $d \times r$ factor loading matrix, each element of which is adapted and has càdlàg paths almost surely. Here, we specifically call F_t the common price factor, in order to distinguish it from the common volatility factor defined later. The common price factor, F_t, and the idiosyncratic component, Z_t, are assumed to follow continuous Itô-semimartingales, and are given by:

$$F_t = F_0 + \int_0^t h_s ds + \int_0^t \eta_s dB_s \tag{35}$$

and

$$Z_t = Z_0 + \int_0^t g_s ds + \int_0^t \gamma_s d\widetilde{B}_s,$$

where B_s and $\widetilde{B}_s$ are independent Brownian motions. All of the coefficient processes, h, η, g and γ are adapted to $(\mathcal{F}_t)_{t \geq 0}$ and have càdlàg paths, almost surely. The above factor model and general settings are discussed in Aït-Sahalia and Xiu (2017).

Now that we have specified a common latent factor model, it remains to take it to the data (i.e., construct volatility forecasts). In order to do this a couple of further ingredients are needed. First, we specify forecasting models. These follow the structure of factor-augmented regressions that are widely exploited in the macroeconomic forecasting literature (see, e.g., Stock and Watson (2002a,b, 2006), Bai and Ng (2006, 2008), and the references cited therein). Namely, consider:

$$y_{t+h} = \alpha' W_t + \beta' \Psi_t + \varepsilon_{t+h}, \tag{36}$$

where y_t is the daily integrated volatility of a target asset being forecasted and h is the forecasting horizon. W_t is a set of observable variables, such as lags of y_t. Ψ_t contains unobservable variables, or usually called the latent factors. In the context of volatility forecasting, we define the latent predictors, Ψ_t, based on the factor structure assumed in (34) and (35):

$$\Psi_t = \int_0^t \text{diag}(\Lambda_s \eta_s \eta_s' \Lambda_s') ds$$

and name it the common volatility factor (also called the IV factors in the sequel) for a natural reason: η_s in the above integrand is the spot volatility of F_t (see (35)). As a result, Ψ_t is defined as the integrated volatility of common price factor F_t. Note that one cannot disentangle Λ from η unless certain identification conditions, such as $\eta\eta' = I_r$, are imposed. However, in the context of forecasting, we don't have to disentangle these components from Ψ_t. This is because we are only interested in Ψ_t, which is the IV matrix of the r uncorrelated common factors in our setup. In summary, it is worth stressing that unlike many other applications of factor-augmented models, we do not directly use weighted common factors, $\Lambda_t F_t$, extracted from a large panel of observable data. Instead, what we actually use as predictors in our forecasting models are the estimated IVs of these common factors (i.e. the Ψ_t).

Of note is that model (36) nests the large family of heterogeneous autoregressive (HAR) type models that are widely used in the literature, including the HAR-RV model discussed earlier. We write a variant of this model as follows:

$$\text{RM}_{t+h} = \alpha_0 + \alpha_1 \text{RM}_t + \alpha_2 \text{RM}_{[t,t-4]} + \alpha_3 \text{RM}_{[t,t-21]} + \epsilon_{t+h}, \tag{37}$$

where RM represents a realized measure of integrated volatility for a target asset, and $\text{RM}_{[t,t-p]}$ is the average of RM's, over the most recent $p+1$ days, i.e. $\text{RM}_{[t,t-p]} = \frac{1}{p+1} \sum_{i=0}^{p} \text{RM}_{t-i}$. As a result, the second and the third variables on the right hand side of (37) represent weekly and the monthly average RM values, respectively. In practice, one might use RV, TRV, or MPV as defined in (31), (32) and (33) as the realized measures in above HAR model. Let $W_t = [1 \ \text{RM}_t \ \text{RM}_{[t,t-4]} \ \text{RM}_{[t,t-21]}]'$ or $W_t = [1 \ \text{RM}_t \ \text{RM}_{[t,t-4]} \ \text{RM}_{[t,t-21]} \ Z_t]'$, where Z_t is a vector of predictors that may be further included in the forecasting model in equation(36).

Heuristically, model (36) considered in this paper combines two distinct sources of information for volatility prediction. The first part, W_t, follows the HAR structure, exploiting time series information on the target asset itself. The second part, Ψ_t, collects further predictive power from broader sources of information, i.e. from a large panel of variables other than the target asset.

Additionally, dimension reduction can be achieved in this setup by simply applying LASSO or elastic

net shrinkage on the IV estimates of all assets in order to first obtain a subset of assets whose IVs are relevant to predicting the target asset's IV. Then, one can apply PCA or sparse PCA (SPCA) to this selected panel of high-frequency asset returns in order to estimate common prices factors, F_t, from which the common IV factors, Ψ_t, are estimated. Finally, these IV factors are incorporated into model (36) to forecast IVs for the target asset.

Recall that the LASSO (see Tibshirani (1996)) and the elastic net (see Zou and Hastie (2005)) can be interpreted as regularized or penalized regression methods. To illustrate, consider a regression of y_{t+h} on W_t and χ_t, where y_{t+h} and W_t are defined in (36), and χ_t is a vector of integrated volatility on day t, for all assets in X_t. The LASSO estimator is the solution to the following problem:

$$\min_{\phi} \sum_{t} \left\{ \left|\left| y_{t+h} - \alpha' W_t - \sum_{j} \phi_j \chi_{j,t} \right|\right|^2 + \lambda \sum_{j} |\phi_j| \right\},$$

where the ϕ's are regression coefficients in a standard penalized regression, and all other parameters are defined above. Similarly, the elastic net estimator is a solution to the following problem:

$$\min_{\phi} \sum_{t} \left\{ \left|\left| y_{t+h} - \alpha' W_t - \sum_{j} \phi_j \chi_{j,t} \right|\right|^2 + \lambda \sum_{j} \left(\frac{(1-\theta)}{2} \phi_j^2 + \theta |\phi_j| \right) \right\}, \tag{38}$$

where $\theta \in [0,1]$. Of note is that when $\theta = 1$, the elastic net is equivalent to the LASSO. Also, as θ shrinks toward 0, elastic net estimators approach those obtained via ridge regression. Furthermore, note that LASSO imposes an $\mathcal{L}_1$-norm penalty on coefficients in the model, while the elastic net induces double shrinkage, in the sense that it imposes a linear combination of $\mathcal{L}_1$-norm and $\mathcal{L}_2$-norm penalties on coefficients in the model. Finally, recall that it is the imposition of the $\mathcal{L}_1$-norm penalty that induces shrinkage to zero of some coefficients in the regression model; and it is the non-zero coefficients in the solution to these minimization problems that are used to select the final set of variables for use when constructing factor estimates. For further discussion of shrinkage methods in economics and finance, see Swanson and Xiong (2018), and for further discussion of the methods discussed here, refer to the paper from which this example is drawn (i.e. Cheng et al. (2021)).

Recall also that PCA and SPCA procedures, are both easy to implement. To start, consider the following covariance matrix estimator, defined on a fixed interval, $[0,T]$:

$$\widehat{\Sigma}_t = \frac{1}{t} \sum_{i=1}^{\lfloor t/\Delta_n \rfloor} \left\{ (\Delta_i^n X)(\Delta_i^n X)' \right\} \mathbf{1}_{\{\|\Delta_i^n X\| \leq c\Delta_n^{\varpi}\}}, \quad \forall t \in [0,T].$$

One carries out PCA by applying an eigenvalue-eigenvector decomposition to $\widehat{\Sigma}_t$, yielding r estimated eigenvalues, in descending order, say $\widehat{\lambda}_1 > \widehat{\lambda}_2 > \cdots > \widehat{\lambda}_r$, and corresponding estimated eigenvectors, $\widehat{\xi}_1, \widehat{\xi}_2, \cdots,$

$\widehat{\xi}_r$. The first r principal components on the fixed interval are estimated as follows:

$$\widehat{\Delta_i^n F}_j = \widehat{\xi}_j' \Big(\Delta_i^n X \Big) \mathbf{1}_{\{\|\Delta_i^n X\| \leq c\Delta_n^{\varpi}\}}, \quad j = 1, \cdots, r.$$

With these estimated principal components, latent common volatility factors on day t can subsequently be estimated as follows:

$$\widehat{\Psi}_{j,t} = \frac{1}{t} \sum_{i=1}^{\lfloor t/\Delta_n \rfloor} \Big(\widehat{\Delta_i^n F}_j \Big)^2, \quad j = 1, \cdots, r.$$

Thus, for any $j = 1, \cdots, r$, we have $\widehat{\Psi}_{j,t} = \widehat{\xi}_j' \widehat{\Sigma}_t \widehat{\xi}_j = \widehat{\lambda}_j \widehat{\xi}_j' \widehat{\xi}_j$, which is equivalent to $\widehat{\lambda}_j$, if the eigenvector has unit-length. In general, PCA yields nonzero factor loadings for (almost) all variables, which exacerbates difficulty in interpretation, and can induce noisiness in estimated factors, especially when ultra high frequency data are used. To avoid these drawbacks, and to induce further parsimony, one can instead utilize SPCA. The key to SPCA is that it yields sparse factor loadings, in the sense that loadings may be identically zero, a feature not feasible in the context of certain types of shrinkage on the $\mathcal{L}_2$-norm, such as that associated with ridge regression.

5. Methods for Evaluating Return and Volatility Forecasts

There is a vast literature discussing methods for evaluating return and volatility forecasts. Of course, if the objective of a practitioner or researcher is to maximize investment performance, then a natural approach is to evaluate the performance of forecasts and to choose the 'best' forecasting model by simply comparing out-of-sample investment performance across models. In this scenario one might estimate in-sample parameters of the models used to construct the forecasts that are in turn used to make investment decisions by using a 'profit' loss function rather than a least squares or maximum likelihood loss function for parameter estimation.

In addition to assessing profits associated with using certain forecasts after implementing a trading rule, say, forecast evaluation can also be done using standard point based measures like mean square forecast errors, mean absolute forecast error deviations, probability scores, contingency tables, or a whole host of other statistics. For a discussion of a few of these, see Stekler (1991) and Stekler (1994). Forecast evaluation can also be done using predictive density and conditional distribution accuracy testing and model selection, such as the methods outlined in Corradi et al. (2009) and Corradi et al. (2011). For a comprehensive discussion of methods for evaluating return and volatility forecasts, see Corradi and Swanson (2006) and all of the other papers in the Handbook of Forecasting series in which this paper is published. In order to illustrate different approaches used when assessing forecasts, we briefly describe

two methods used for predictive accuracy testing and model selection.

First, consider the so-called Diebold-Mariano test. In the context of forecast model selection, Diebold and Mariano (1995) propose a test for the null hypothesis of equal predictive ability. In its simplest form, the Diebold and Mariano test allows for nondifferentiable loss functions, but does not explicitly account for parameter estimation error, instead relying on the assumption that the in-sample estimation period grows more quickly than the out-of-sample prediction period, so that parameter estimation error vanishes asymptotically. West (1996) takes the more general approach of explicitly allowing for parameter estimation error, although at the cost of assuming that the loss function used in test statistic construction is differentiable. Let $u_{0,t+h}$ and $u_{1,t+h}$ be the $h-$step ahead prediction error associated with predictions of y_{t+h}, using information available up to time t. For example, for $h = 1$, $u_{0,t+1} = y_{t+1} - \kappa_0(Z_0^{t-1}, \theta_0^{\dagger})$, and $u_{1,t+1} = y_{t+1} - \kappa_1(Z_1^{t-1}, \theta_1^{\dagger})$, where Z_0^{t-1} and Z_1^{t-1} contain past values of y_t and possibly other conditioning variables, and where the κ are functions that may be linear, for example. Assume that the two models are nonnested (i.e. Z_0^{t-1} not a subset of Z_1^{t-1}, and vice-versa and/or the function $\kappa_1 \neq \kappa_0$). As is well known, the ranking of models based on their predictive accuracy depends on the loss function used, under generic misspecification (i.e., for the case where all models are approximations of the true underlying model). Hereafter, denote the loss function as $g(\cdot)$, and let $T = R + P$, where only the last P observations are used for model evaluation. Under the assumption that $u_{0,t}$ and $u_{1,t}$ are strictly stationary, the null hypothesis of equal predictive accuracy is specified as:

$$H_0 : E(g(u_{0,t}) - g(u_{1t})) = 0$$

and

$$H_A : E(g(u_{0,t}) - g(u_{1t})) \neq 0$$

In practice, we do not observe $u_{0,t+1}$ and $u_{1,t+1}$, but only $\widehat{u}_{0,t+1}$ and $\widehat{u}_{1,t+1}$, where $\widehat{u}_{i,t+1} = y_{t+1} - \kappa_0(Z_0^t, \widehat{\theta}_{0,t})$, and where $\widehat{\theta}_{i,t}$ is an estimator constructed using observations from 1 up to t, $t \geq R$, in the recursive window model estimation estimation case, and between $t - R + 1$ and t in the rolling window model estimation case, for $i = 1, 2$. For brevity, we consider the recursive scheme. Note that the rolling scheme can be treated in an analogous manner. Of crucial importance is the loss function used for estimation. In fact, if we use the same loss function for estimation and model evaluation, the contribution of parameter estimation error is asymptotically negligible, regardless of the limit of the ratio P/R as $T \to \infty$. Here, for $i = 0, 1$, we set:

$$\widehat{\theta}_{i,t} = \arg\min_{\theta_i \in \Theta_i} \frac{1}{t} \sum_{j=1}^{t} q(y_j - \kappa_i(Z_i^{j-1}, \theta_i)), \ t \geq R$$

In the sequel, we assume that g is continuously differentiable. The case of non-differentiable loss functions is treated by McCracken (2004). The Diebold and Mariano statistic that one uses in this framework to test the above hypotheses is:

$$DM_P = \frac{1}{\sqrt{P}} \frac{1}{\widehat{\sigma}_P} \sum_{t=R}^{T-1} \left(g(\widehat{u}_{0,t+1}) - g(\widehat{u}_{1,t+1}) \right).$$

This statistic has a standard normal limiting distribution under the null hypothesis when $q(\cdot) = g(\cdot)$ or when parameter estimation error vanishes asymptotically, as discussed in Corradi and Swanson (2006).[26]

The above discussion is for the case of one-step ahead prediction errors. All results carry over to the case of $h > 1$. However, in the multistep ahead case, one needs to decide whether to compute 'direct' $h-$step ahead forecast errors (i.e. $\widehat{u}_{i,t+h} = y_{t+h} - \kappa_i(Z_i^{t-h}, \widehat{\theta}_{i,t})$) or to compute iterated $h-$ahead forecast errors (i.e. first predict y_{t+1} using observations up to time t, and then use this predicted value in order to predict y_{t+2}, and so on).

Now, consider the reality check or data snooping test due to White (2000). White proposes an approach for choosing amongst many different models. Suppose there are m models, and we select model 1 as our benchmark (or reference) model. Models $i = 2, ..., m$ are called the competitor (alternative) models. Typically, the benchmark model is either a simple model, our favorite model, or the most commonly used model. Given the benchmark model, the objective is to answer the following question: Is there any model, amongst the set of $m-1$ competitor models, that yields more accurate predictions (for the variable of interest) than the benchmark? As above, let the forecast error be $u_{i,t+1} = y_{t+1} - \kappa_i(Z^t, \theta_i^\dagger)$, and let $\widehat{u}_{i,t+1} = y_{t+1} - \kappa_i(Z^t, \widehat{\theta}_{i,t})$, where $\kappa_i(Z^t, \widehat{\theta}_{i,t})$ is the conditional mean function under model i, and $\widehat{\theta}_{i,t}$ is defined as above. Assume that the set of regressors may vary across different models, so that Z^t is meant to denote the collection of all potential regressors. Define the statistic

$$S_P = \max_{k=2,...,m} S_P(1,k),$$

where

$$S_P(1,k) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} \left(g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1}) \right), \ k = 2, ..., m,$$

---

[26] In this setup, recall that if $g(\cdot)$ is the quadratic loss function, then:

$$g(\widehat{u}_{0,t}) - g(\widehat{u}_{1,t}) = \widehat{u}_{0,t}^2 - \widehat{u}_{1,t}^2.$$

This example corresponds to the case where the mean square forecast error of the two competing models is being compared.

Thus, our statistic is simply the "max" of all of the pairwise DM test statistics constructed using model 1 as the "benchmark".

The test hypotheses are formulated as

$$H_0 : \max_{k=2,\ldots,m} E(g(u_{1,t+1}) - g(g_{k,t+1})) \leq 0$$

$$H_A : \max_{k=2,\ldots,m} E(g(u_{1,t+1}) - g(u_{k,t+1})) > 0,$$

where $u_{k,t+1} = y_{t+1} - \kappa_k(Z^t, \theta^\dagger_{k,t})$, and $\theta^\dagger_{k,t}$ denotes the probability limit of $\theta_{i,t}$. Thus, under the null hypothesis, no competitor model, amongst the set of the $m-1$ alternatives, can provide a more (loss function specific) accurate prediction than the benchmark model. On the other hand, under the alternative, at least one competitor (and in particular, the best competitor) provides more accurate predictions than the benchmark. Given nonestedness, White shows that under H_0, $\max_{k=2,\ldots,m} S_P(1,k)$ is a zero mean Gaussian process. Evidently, the White statistic is a 'max' statistic which is simply the maximum of m pairwise Diebold Mariano statistics calculated by comparing each competitor model with Model 1. Valid critical values for use in implementation of this test can be constructed quite easily using the block bootstrap, as discussed in Corradi and Swanson (2006). For a detailed discussion of this test, refer to Corradi and Swanson (2007).

5.1. Methods for Evaluating Investments

Ultimately, the 'proof is in the pudding' when it comes to evaluating investment performance. Namely, how much wealth is generated, and how much risk is taken under alternative investment scenarios. If all that one cares about is wealth, in the sense that any measure of risk is tolerable, then one need simply look at the dollar value of alternative investments made using predictions from alternative forecasting models and methods. If one is interested in measuring investment performance based on an assessment of both mean and variance considerations (i.e., by considering both accumulated wealth and the risk associated with the investment), then mean-variance trade-offs are important, as discussed above. As discussed in Makridakis et al. (2024), the investment component of the M6 competition is decided by using a variant of the well known Sharp and Information Ratios. The particular statistic used in their paper is called IR, and is defined to be the ratio of portfolio returns ret to the standard deviation of portfolio returns, sdp, or $IR = \frac{ret}{sdp}$ When applying these sorts of measures, one must choose the period over which returns and risk are calculated. Needless to say, many variants of the above measure as well as other related measures are widely used by practitioners and researchers. Moreover, the particular measure used is usually determined on a case by case basis depending on the stated objectives or goals

of the investor. For example, risk tolerance may play an important role in portfolio optimization. The performance measure used to evaluate a portfolio where particularly risky stocks are not included may be quite different from that used to evaluate portfolios that include only risk free assets such as U.S. treasuries.

It is also worth noting that approaches used for calibrating and estimating forecasting models can be connected in a natural way to investment performance evaluation. In particular, note that many forecasting models of the variety discussed above are estimated using standard statistical methods such as maximum likelihood, simulated maximum likelihood, and methods of moments. Take the case of maximum likelihood (ML). In its simplest form, given a number of assumptions, ML equates with least squares, which is an estimator often used to obtain coefficient estimates of linear models. However, one might ask why least squares is used to calibrate forecasting models when the ultimate objective is profit maximization, say. For example, assume that the investment objective is simply to maximize dollar wealth based on a given portfolio. In that case, rather than estimating model parameters using least squares, one might estimate them using an objective function that is based on the performance of the model when constructing forecasts for use in building an investment portfolio. Namely, directly calibrate the forecasting model to maximize investment returns. Of course, if one estimates forecasting models using an estimator other than least squares or nonlinear least squares, say, certain optimality properties associated with the least squares estimator might be lost. That said, one might not care about such issues if the overarching objective of an exercise is to maximize returns.

6. Empirical Illustration: Forecasting Returns Using Latent Factor and Variable Selection Methods

In this illustration, we construct daily return forecasts for 391 constituents of the S&P500 index at h =1-day, 5-day, and 20-day ahead prediction horizons. All data are collected from the well-known TAQ database. The 391 stocks for which we collect price and return data include the constituents of the S&P 500 index that were in the index for our entire sample period of February 28, 2009 - January 28, 2019 (2533 trading days). In our analysis, all forecasting models are estimated using rolling (fixed) windows of 1259 days.

The forecasting models that we evaluate are summarized in Table 1, and all have the following functional form:

$$r_{t+h} = \alpha + I\{p > 0\}(\sum_{i=1}^{p} \beta_i r_{t-i+1}) + \gamma' F_t + \epsilon_{t+h}, \tag{39}$$

where r_{t+h}, is the log return for a given stock, F_t is an $r-$dimensional vector of estimated factors, constructed using (sparse) principal components analysis (PCA), ϵ_{t+h} is a stochastic disturbance term, and $I\{\cdot\}$ denotes the indicator function, which takes the value one if a nonzero number of lags selected, using the Schwarz Information Criterion (SIC), and zero otherwise. Note that all models and lag orders are selected individually for each stock prior to the construction of each new prediction, at each point in time as we iterate through the ex-ante sample period. Our benchmark model sets $\gamma = 0$, so that no factors are included in the forecasting equation. Factors are estimated using variable selection and dimension reduction methods. These include: AR+PCA, where factors are estimated using PCA applied to our entire return dataset; AR+SPCA, where factors are estimated using SPCA applied to our entire return dataset; and AR+HT, where factors are estimated using PCA applied to a subset of our return variable dataset that is selected using hard thresholding (HT), as discussed in Bai and Ng (2008). Specifically, for our AR+HT model, for each variable in our dataset, say $\mathbf{X}$, and for each forecast horizon, h, we perform regressions of r_{t+h} on lags of r_t and on $X_{i,t}$, where $X_{i,t}$ is a scalar variable in $\mathbf{X}$, for $i = 1, ..., 390$, and lags of r_t are selected using the SIC. Finally, let t_i denote the t statistic associated with X_{it-h} in the regression, and select variable X_{it} if $|t_i| > 1.28$. Then, utilize PCA to estimate factors for inclusion in the above forecasting equation, given the set of variables selected after iterating through all candidate variables in $\mathbf{X}$. Should less than 20 variables be selected, use the AR(SIC) model. As models are re-estimated at each point in time, this approach is a hybrid approach, in the sense that some models may include factors as regressors, while others may be simple AR(SIC) models. Note that in our experiments, less than 10% of the total number of forecasting periods involved replacing the AR+HT model with our AR(SIC) benchmark.

Results based on the above experiment are presented for three different forecasting periods, including: February 28, 2014 - Februray 25, 2016 (Subsample1), Februray 25, 2016 - January 24, 2019 (Subsample2), and February 28, 2014 - January 24, 2019 (Full Sample). In addition to evaluating the performance our our 4 individual models (i.e., the benchmark AR(SIC), PCA, SPCA, and AR-HT models, we evaluate the performance of two different model combinations in which forecasts are averaged, including: Combination 1 – AR+PCA and AR+SPCA, and Combination 2 – AR+PCA and AR+SPCA and AR+HT. All reported results are based on mean square forecast errors calculated across a given sample period, for a given variable and forecast horizon. Our models are summarized in Table 1, while the forecast combinations analyzed are given in Table 2. Tables 3 and 4 summarize results based on the comparison of our individual models, while Tables 5 and 6 present summary findings based on forecasts constructed using our forecast

combinations. Our findings can be summarized as follows.

Turning first to Table 3, notice that the AR+HT model clearly performs 'best' for the majority of stocks in our sample, regardless of sample period or forecast horizon. Moreover, the pure AR benchmark clearly performs worst, as can be seen by noting the 391 minus the sum of all 'wins' across a given row of entries in this table gives the number of stocks for which the AR models 'wins'. This finding indicates that more complicated forecasting models outperform simple AR (or random walk) type models, as claimed by Lo and MacKinlay (1999). Second, note that the 'average ranking' of the AR+HT model is always lower than that of any other model, as can be seen by examining the entries in Table 4. This result supports our findings from Table 3. Third, forecast combination yields substantially more 'wins' than the very best of our 4 'non-combination' models, as can be seen upon inspection of the results presented in Table 5. Finally, the results in Table 6 indicate that the 'average ranking' of Combination 2 (i.e., our best combination method) is as close to a perfect ranking of '1' as is the 'average ranking' of AR-HT when the latter model is compared solely with our other three non-combination models.

7. Closing Remarks

In this paper we discuss various developments in modelling and forecasting financial time series models that have been made over the last 25 years, for use when analyzing small,large, and very large datasets. Models discussed range from discrete GARCH models to machine learning models based on the least absolute shrinkage and selection operator and the elastic net. We also briefly discuss earlier and simpler methods ranging from the use of surveys to the specification of simple autoregressive regression models. Finally, we present the results of a small empirical illustration in which returns for a number of stocks are predicted using simple AR type models as well various machine learning type models.

References

Aït-Sahalia, Y., Jacod, J., 2014. High-frequency financial econometrics. Princeton University Press, Princeton.

Aït-Sahalia, Y., Jacod, J., et al., 2009. Testing for jumps in a discretely observed process. The Annals of Statistics 37, 184–222.

Aït-Sahalia, Y., Xiu, D., 2017. Using principal component analysis to estimate a high dimensional factor model with high-frequency data. Journal of Econometrics 201, 384–399.

Alexander, C., Leigh, C.T., 1997. On the covariance matrices used in value at risk models .

Andersen, T., Bollerslev, T., Diebold, F., 2007. Roughing it up: including jump components in the measurement, modeling, and forecasting of return volatility. The Review of Economics and Statistics 89, 701–720.

Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2001. The distribution of realized exchange rate volatility. Journal of the American Statistical Association 96, 42–55.

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

Andersen, T.G., Dobrev, D., Schaumburg, E., 2012. Jump-robust volatility estimation using nearest neighbor truncation. Journal of Econometrics 169, 75–93.

Andersen, T.G., Lund, J., 1997. Estimating continuous-time stochastic volatility models of the short-term interest rate. Journal of Econometrics 77, 343–377.

Arif, S., Lee, C.M., 2014. Aggregate investment and investor sentiment. The Review of Financial Studies 27, 3241–3279.

Bai, J., Li, K., 2012. Statistical analysis of factor models of high dimension. The Annals of Statistics 40, 436 – 465.

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

Bai, J., Ng, S., 2008. Forecasting economic time series using targeted predictors. Journal of Econometrics 146, 304–317.

Baillie, R.T., 1996. Long memory processes and fractional integration in econometrics. Journal of Econometrics 73, 5–59.

Baker, M., Wurgler, J., 2000. The equity share in new issues and aggregate stock returns. The Journal of Finance 55, 2219–2257.

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. The Journal of Finance 61, 1645–1680.

Baker, M., Wurgler, J., 2007. Investor sentiment in the stock market. Journal of Economic Perspectives 21, 129–151.

Barndorff-Nielsen, O.E., Graversen, S.E., Jacod, J., Podolskij, M., Shephard, N., 2006. A central limit theorem for realised power and bipower variations of continuous semimartingales, in: From Stochastic Calculus to Mathematical Finance. Springer, pp. 33–68.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2008. Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. Econometrica 76, 1481–1536.

Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2011. Subsampling realised kernels. Journal of Econometrics 160, 204–219.

Barndorff-Nielsen, O.E., Shephard, N., 2003. Realized power variation and stochastic volatility models. Bernoulli 9, 243–265.

Barndorff-Nielsen, O.E., Shephard, N., 2004. Power and bipower variation with stochastic volatility and jumps. Journal of Financial Econometrics 2, 1–37.

Bartholdy, J., Peare, P., 2003. Unbiased estimation of expected return using capm. International Review of Financial Analysis 12, 69–81.

Basak, S., Kar, S., Saha, S., Khaidem, L., Dey, S.R., 2019. Predicting the direction of stock market prices using tree-based classifiers. The North American Journal of Economics and Finance , 552–567.

Bates, D.S., 2000. Post-'87 crash fears in the s&p 500 futures option market. Journal of Econometrics 94, 181–238.

Bates, J.M., Granger, C.W., 1969. The combination of forecasts. Journal of the Operational Research Society 20, 451–468.

Bauwens, L., Chevillon, G., Laurent, S., 2023. We modeled long memory with just one lag! Journal of Econometrics 236, 105467.

Black, F., 1976. Studies in stock price volatility changes, in: Proceedings of the 1976 Meetings of the Business and Economic Statistics Section, American Statistical Association. pp. 177–181.

Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. Journal of Political Economy 81, 637–654.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.

Bollerslev, T., 2008. Glossary to arch (garch). CREATES and National Bureau of Economic Research Working Paper No. 49, Duke University .

Bollerslev, T., Engle, R.F., Wooldridge, J.M., 1988. A capital asset pricing model with time-varying covariances. Journal of Political Economy 96, 116–131.

Bollerslev, T., Tauchen, G., Zhou, H., 2009. Expected stock returns and variance risk premia. The Review of Financial Studies 22, 4463–4492.

Brownlees, C.T., Nualart, E., Sun, Y., 2016. On the estimation of integrated volatility in the presence of jumps and microstructure noise. Econometric Reviews 38, 991–1013.

Cai, Z., 2007. Trending time-varying coefficient time series models with serially correlated errors. Journal of Econometrics 136, 163–188.

Çakmaklı, C., van Dijk, D., 2016. Getting the most out of macroeconomic information for predicting excess stock returns. International Journal of Forecasting 32, 650–668.

Campbell, J.Y., 1987. Stock returns and the term structure. Journal of financial economics 18, 373–399.

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

Chacko, G., Viceira, L.M., 2003. Spectral gmm estimation of continuous-time processes. Journal of Econometrics 116, 259–292.

Chao, J.C., Qiu, K., Swanson, N.R., Liu, Y., 2024. A completely consistent variable selection method for forecasting with factor models. Rutgers University Working Paper .

Chen, B., Maung, K., 2023. Time-varying forecast combination for high-dimensional data. Journal of Econometrics 237, 405–418.

Chen, L., Pelger, M., Zhu, J., 2024. Deep learning in asset pricing. Management Science 70, 714–750.

Cheng, M., Swanson, N.R., Yang, X., 2021. Forecasting volatility using double shrinkage methods. The Journal of Empirical Finance 62, 46–61.

Chernov, M., Ghysels, E., 2000. A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of options valuation. Journal of Financial Economics 56, 407–458.

Chincarini, Ludwig, B., Kim, D., 2006. Quantitative Equity Portfolio Management. McGraw Hill, New York.

Christie, A.A., 1982. The stochastic behavior of common stock variances: Value, leverage and interest rate effects. Journal of Financial Economics 10, 407–432.

Clements, M.P., Rich, R.W., Tracy, J.S., 2023. Surveys of professionals, in: Handbook of Economic Expectations. Elsevier, pp. 71–106.

Cochrane, J.H., 1991. Production-based asset pricing and the link between stock returns and economic fluctuations. The Journal of Finance 46, 209–237.

Corradi, V., Distaso, W., Swanson, N.R., 2009. Predictive density estimators for daily volatility based on the use of realized measures. Journal of Econometrics 150, 119–138.

Corradi, V., Distaso, W., Swanson, N.R., 2011. Predictive inference for integrated volatility. Journal of the American Statistical Association 106, 1496–1512.

Corradi, V., Swanson, N.R., 2006. Predictive density evaluation, in: Handbook of Economic Forecasting, Volume 1. Elsevier, pp. 197–284.

Corradi, V., Swanson, N.R., 2007. Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes. International Economic Review 48, 67–109.

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

Corsi, F., Pirino, D., Reno, R., 2010. Threshold bipower variation and the impact of jumps on volatility forecasting. Journal of Econometrics 159, 276–288.

Cox, J.C., Ingersoll, J.E.J., Ross, S.A., 1985. A theory of the term structure of interest rates. Econometrica 53, 385–407.

Cox, J.C., Ross, S.A., 1976. The valuation of options for alternative stochastic processes. Journal of Financial Economics 3, 145–166.

Dangl, T., Halling, M., 2012. Predictive regressions with time-varying coefficients. Journal of Financial Economics 106, 157–181.

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

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

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica , 987–1007.

Engle, R.F., Bollerslev, T., 1986. Modelling the persistence of conditional variances. Econometric Reviews 5, 1–50.

Engle, R.F., Gonzalez-Rivera, G., 1991. Semiparametric arch models. Journal of Business & Economic Statistics 9, 345–359.

Engle, R.F., Lilien, D.M., Robins, R.P., 1987. Estimating time varying risk premia in the term structure: The arch-m model. Econometrica , 391–407.

Fabozzi, F.J., Gupta, F., Markowitz, H.M., 2002. The legacy of modern portfolio theory. The Journal of Investing 11, 7–22.

Fama, E.F., 1965. The behavior of stock-market prices. The Journal of Business 38, 34–105.

Fama, E.F., 1995. Random walks in stock market prices. Financial Analysts Journal 51, 75–80.

Fama, E.F., MacBeth, J.D., 1973. Risk, return, and equilibrium: Empirical tests. Journal of Political Economy 81, 607–636.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

Fan, J., Wang, Y., 2007. Multi-scale jump and volatility analysis for high-frequency financial data. Journal of the American Statistical Association 102, 1349–1362.

Farmer, L.E., Schmidt, L., Timmermann, A., 2023. Pockets of predictability. The Journal of Finance 78, 1279–1341.

French, K.R., Schwert, G.W., Stambaugh, R.F., 1987. Expected stock returns and volatility. Journal of Financial Economics 19, 3–29.

Frydman, R., Mangee, N., Stillwagon, J., 2021. How market sentiment drives forecasts of stock returns. Journal of Behavioral Finance 22, 351–367.

Glosten, L.R., Jagannathan, R., Runkle, D.E., 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. The Journal of Finance 48, 1779–1801.

Goyal, A., Welch, I., 2003. Predicting the equity premium with dividend ratios. Management Science 49, 639–654.

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

Goyal, A., Welch, I., Zafirov, A., 2023. A comprehensive 2022 look at the empirical performance of equity premium prediction. Swiss Finance Institute Research Paper .

Granger, C.W., Ramanathan, R., 1984. Improved methods of combining forecasts. Journal of Forecasting 3, 197–204.

Groen, J.J., Kapetanios, G., 2016. Revisiting useful approaches to data-rich macroeconomic forecasting. Computational Statistics & Data Analysis 100, 221–239.

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

Hamilton, J.D., 2016. Macroeconomic regimes and regime shifts. Handbook of Macroeconomics 2, 163–201.

Harvey, A., Ruiz, E., Shephard, N., 1994. Multivariate stochastic variance models. The Review of Economic Studies 61, 247–264.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical learning with sparsity. Monographs on Statistics and Applied Probability 143, 8.

Hautsch, N., Podolskij, M., 2013. Preaveraging-based estimation of quadratic variation in the presence of noise and jumps: theory, implementation, and empirical evidence. Journal of Business & Economic Statistics 31, 165–183.

Henkel, S.J., Martin, J.S., Nardari, F., 2011. Time-varying short-horizon predictability. Journal of Financial Economics 99, 560–580.

Heston, S.L., 1993. A closed-form solution for options with stochastic volatility with applications to bond and currency options. The Review of Financial Studies 6, 327–343.

Hirshleifer, D., Hou, K., Teoh, S.H., 2009. Accruals, cash flows, and aggregate stock returns. Journal of Financial Economics 91, 389–406.

Hsieh, D.A., 1991. Chaos and nonlinear dynamics: application to financial markets. The Journal of Finance 46, 1839–1877.

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

Jacod, J., Li, Y., Zheng, X., 2019. Estimating the integrated volatility with tick observations. Journal of Econometrics 37, 80–100.

Jacod, J., Rosenbaum, M., et al., 2013. Quarticity and other functionals of volatility: efficient estimation. The Annals of Statistics 41, 1462–1484.

Jacod, J., Todorov, V., et al., 2014. Efficient estimation of integrated volatility in presence of infinite variation jumps. The Annals of Statistics 42, 1029–1069.

Jiang, J., Kelly, B., Xiu, D., 2023. (re-) imag (in) ing price trends. The Journal of Finance 78, 3193–3249.

Jing, B.Y., Liu, Z., Kong, X.B., 2014. On the estimation of integrated volatility with jumps and microstructure noise. Journal of Business & Economic Statistics 32, 457–467.

Kabiri, A., Landon-Lane, J., Qiu, K., Swanson, N.R., Turton, J., Tuckett, D., 2023. The impact of sentiment on judgement and econometric model based forecasts. Rutgers University Working Paper .

Kalnina, I., Linton, O., 2008. Estimating quadratic variation consistently in the presence of endogenous and diurnal measurement error. Journal of Econometrics 147, 47–59.

Kelly, B., Pruitt, S., 2013. Market expectations in the cross-section of present values. The Journal of Finance 68, 1721–1756.

Kelly, B., Pruitt, S., 2015. The three-pass regression filter: A new approach to forecasting using many predictors. Journal of Econometrics 186, 294–316.

Lawrence, M., Goodwin, P., O'Connor, M., Önkal, D., 2006. Judgmental forecasting: A review of progress over the last 25 years. International Journal of Forecasting 22, 493–518.

Lee, J.H., Shi, Z., Gao, Z., 2022. On lasso for predictive regression. Journal of Econometrics 229, 322–349.

Lettau, M., Ludvigson, S., 2001. Consumption, aggregate wealth, and expected stock returns. The Journal of Finance 56, 815–849.

Li, H., Hong, Y., 2011. Financial volatility forecasting with range-based autoregressive volatility model. Finance Research Letters 8, 69–76.

Lin, H., Wu, C., Zhou, G., 2018. Forecasting corporate bond returns with a large set of predictors: An iterated combination approach. Management Science 64, 4218–4238.

Lo, A.W., MacKinlay, C., 1999. A Non-Random Walk Down Wall Street. Princeton University Press, Princeton.

Lo, A.W., Mamaysky, H., Wang, J., 2000. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. The Journal of Finance 55, 1705–1765.

Long, S., Lucey, B., Xie, Y., Yarovaya, L., 2023. "i just like the stock": The role of reddit sentiment in the gamestop share rally. Financial Review 58, 19–37.

Ludvigson, S.C., Ng, S., 2007. The empirical risk–return relation: A factor analysis approach. Journal of Financial Economics 83, 171–222.

Makridakis, S., Spiliotis, E., Hollyman, R., Petropoulos, F., Swanson, N.R., Gaba, A., 2024. The M6 forecasting competition: Bridging the gap between forecasting and investment decisions. arXiv preprint arXiv:2310.13357 .

Mancini, C., 2001. Disentangling the jumps of the diffusion in a geometric jumping brownian motion. Giornale dell'Istituto Italiano degli Attuari 64, 44.

Mancini, C., 2009. Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. Scandinavian Journal of Statistics 36, 270–296.

Mandelbrot, B., 1963. The variation of certain speculative prices. The Journal of Business 36, 394–419.

Markowitz, H.M., 1952. Portfolio selection. The Journal of Finance 7, 77–91.

McCracken, M.W., 2004. Parameter estimation error and tests of equal forecast accuracy between non-nested models. International Journal of Forecasting 20, 503–514.

McCracken, M.W., 2007. Asymptotics for out of sample tests of granger causality. Journal of Econometrics 140, 719–752.

McMillan, D.G., Kambouroudis, D., 2009. Are riskmetrics forecasts good enough? evidence from 31 stock markets. International Review of Financial Analysis 18, 117–124.

Medeiros, M.C., Vasconcelos, G.F., Veiga, Á., Zilberman, E., 2021. Forecasting inflation in a data-rich environment: the benefits of machine learning methods. Journal of Business & Economic Statistics 39, 98–119.

Merton, R.C., 1976. Option pricing when underlying stock returns are discontinuous. Journal of Financial Economics 3, 125–144.

Miccolis, J.A., Goodman, M., 2012. Next generation investment risk management: Putting the'modern'back in modern portfolio theory. Journal of Financial Planning 25.

Mina, J., Xiao, J.Y., et al., 2001. Return to riskmetrics: the evolution of a standard. RiskMetrics Group 1, 1–11.

Møller, S.V., Rangvid, J., 2015. End-of-the-year economic growth and time-varying expected returns. Journal of Financial Economics 115, 136–154.

Mukherjee, A., Peng, W., Swanson, N.R., Yang, X., 2020. Financial econometrics and big data: A survey of volatility estimators and tests for the presence of jumps and co-jumps, in: Rao, C., Vinod, H. (Eds.), Handbook of Statistics Vol. 42, pp. 179–233.

Mykland, P.A., Zhang, L., 2009. Inference for continuous semimartingales observed at high frequency. Econometrica 77, 1403–1445.

Neely, C.J., Rapach, D.E., Tu, J., Zhou, G., 2014. Forecasting the equity risk premium: the role of technical indicators. Management Science 60, 1772–1791.

Nelson, D.B., 1990. Arch models as diffusion approximations. Journal of Econometrics 45, 7–38.

Nelson, D.B., 1991. Conditional heteroskedasticity in asset returns: A new approach. Econometrica , 347–370.

Pan, J., 2002. The jump-risk premia implicit in options: Evidence from an integrated time-series study. Journal of Financial Economics 63, 3–50.

Pesaran, M.H., Timmermann, A., 1995. Predictability of stock returns: Robustness and economic significance. The Journal of Finance 50, 1201–1228.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., Siemsen, E., 2018. Judgmental selection of forecasting models. Journal of Operations Management 60, 34–46.

Pettenuzzo, D., Timmermann, A., Valkanov, R., 2014. Forecasting stock returns under economic constraints. Journal of Financial Economics 114, 517–553.

Podolskij, M., Vetter, M., et al., 2009. Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps. Bernoulli 15, 634–658.

Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. Journal of the American Statistical Association 89, 1303–1313.

Poskitt, D.S., 2007. Autoregressive approximation in nonstandard situations: The fractionally integrated and non-invertible cases. Annals of the Institute of Statistical Mathematics 59, 697–725.

Poterba, J.M., Summers, L.H., 1986. Reporting errors and labor market dynamics. Econometrica , 1319–1338.

Rapach, D., Zhou, G., 2013. Forecasting stock returns, in: Handbook of Economic Forecasting. Elsevier. volume 2, pp. 328–383.

Rapach, D.E., Ringgenberg, M.C., Zhou, G., 2016. Short interest and aggregate stock returns. Journal of Financial Economics 121, 46–65.

Rapach, D.E., Strauss, J.K., Zhou, G., 2010. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. The Review of Financial Studies 23, 821–862.

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

Rasmussen, M., 2003. Quantitative Portfolio Optimisation, Asset Allocation and Risk Management. MacMillan, New York.

Robinson, P.M., 1989. Nonparametric Estimation of Time-varying Parameters. Springer, New York.

Rossi, A.G., 2018. Predicting stock market returns with machine learning. Uiversity of Maryland Working Paper .

Schwert, G.W., 1989. Why does stock market volatility change over time? The Journal of Finance 44, 1115–1153.

Schwert, G.W., 1990. Stock volatility and the crash of'87. The Review of Financial Studies 3, 77–102.

Stark, T., 2013. Spf panelists' forecasting methods: A note on the aggregate results of a november 2009 special survey. Federal Reserve Bank of Philadelphia Working Paper .

Stekler, H.O., 1991. Macroeconomic forecast evaluation techniques. International Journal of Forecasting 7, 375–384.

Stekler, H.O., 1994. Are economic forecasts valuable. Journal of Forecasting 13, 495–505.

Stock, J.H., Watson, M.W., 2002a. Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97, 1167–1179.

Stock, J.H., Watson, M.W., 2002b. Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics 20, 147–162.

Stock, J.H., Watson, M.W., 2004. Combination forecasts of output growth in a seven-country data set. Journal of Forecasting 23, 405–430.

Stock, J.H., Watson, M.W., 2006. Forecasting with many predictors. Handbook of Economic Forecasting 1, 515–554.

Stock, J.H., Watson, M.W., 2016. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics, in: Handbook of Macroeconomics. Elsevier. volume 2, pp. 415–525.

Su, L., Wang, X., 2017. On time-varying factor models: Estimation and testing. Journal of Econometrics 198, 84–101.

Swanson, N.R., Xiong, W., 2018. Big data analytics in economics: What have we learned so far, and where should we go from here? Canadian Journal of Economics 51, 695–746.

Taylor, S.J., 1986. Modelling Financial Time Series. John Wiley and Sons, New York.

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

Timmermann, A., 2006. Forecast combinations. Handbook of Economic Forecasting 1, 135–196.

Timmermann, A., 2008. Elusive return predictability. International Journal of Forecasting 24, 1–18.

Todorov, V., Tauchen, G., 2012. The realized laplace transform of volatility. Econometrica 80, 1105–1127.

de Vincent-Humphreys, R., Dimitrova, I., Falck, E., Henkel, L., Meyler, A., 2019. Twenty years of the ecb survey of professional forecasters. ECB Economic Bulletin Articles 1, 1–35.

Wang, C.S., Wan, S.K., 2020. A var approach to forecasting multivariate long memory processes subject to structural breaks, in: Essays in Honor of Cheng Hsiao. Emerald Publishing Limited, pp. 105–141.

Wang, C.S.H., Bauwens, L., Hsiao, C., 2013. Forecasting a long memory process subject to structural breaks. Journal of Econometrics 177, 171–184.

West, K.D., 1996. Asymptotic inference about predictive ability. Econometrica 64, 1067–1084.

White, H., 2000. A reality check for data snooping. Econometrica 68, 1097–1126.

Wold, H., 1966. Estimation of principal components and related models by iterative least squares. Multivariate Analysis , 391–420.

Zhang, L., Mykland, P.A., Aït-Sahalia, Y., 2005. A tale of two time scales: Determining integrated volatility with noisy high-frequency data. Journal of the American Statistical Association 100, 1394–1411.

Zhang, L., et al., 2006. Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach. Bernoulli 12, 1019–1043.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 67, 301–320.

Table 1: Models Used in Prediction Experiments*

Model	Description
Model 1: AR(SIC)/RW	Autoregressive model with lag(s) selected by the Schwarz information criterion/ Random walk model.
Model 2: AR+PCA	AR(SIC) model augmented with PCA type factors constructed using our entire return dataset.
Model 3: AR+SPCA	AR(SIC) model augmented with SPCA type factors constructed using our entire return dataset.
Model 4: AR+HT	AR(SIC) model augmented with HT type factors constructed using PCA applied to a subset of our return dataset that is selected using the hard thresholding method.

*Notes: This table includes brief descriptions of the 4 forecasting models used in daily predictions of returns. For complete details refer to Section 6.

Table 2: Combination Methods Used in Prediction Experiments*

Model	Description
Combination 1	Average of two single models including AR+PCA, AR+SPCA.
Combination 2	Average of all three single models including AR+PCA, AR+SPCA, AR+HT.

*Notes: Entries in this table describe the forecast combinations used in our forecast combination experiments.

Table 3: Comparison of Forecasting Models by Number of Stocks with Lowest MSFE Across Forecast Horizons and Subsamples*

		AR+PCA	AR+SPCA	AR+HT
Full Sample	h=1	46	78	**267**
(2014-02-28 - 2019-01-24)	h=5	54	71	**266**
	h=20	76	53	**262**
Subsample 1	h=1	47	131	**213**
(2014-02-28 - 2016-02-25)	h=5	63	64	**264**
	h=20	73	46	**272**
Subsample 2	h=1	72	54	**265**
(2016-02-26 - 2019-01-24)	h=5	57	84	**250**
	h=20	85	68	**238**

*Notes: See notes to Tables 1 and 2. Entries in this tables are the number of stocks (out of a total of 361) for which each forecasting model (AR-PCA, AR+SPCA, AR+HT) has the lowest MSFE at forecast horizons of 1, 5, and 20 days ahead, across the full sample and two subsamples. Note that as there are 391 stocks in total, when summing entries across rows yields a value of less than 391, this indicates that the pure AR benchmark model "wins" for a number of stocks. The entire sample period is from 2014-02-28 to 2019-01-24, with the first subsample from 2014-02-28 to 2016-02-25, and the second subsample from 2016-02-25 to 2019-01-24. For complete details, refer to Section 6.

Table 4: Table of Average Rankings of All Single Models Across Forecast Horizons and Subsamples*

		AR+PCA	AR+SPCA	AR+HT
Full Sample	h=1	2.28	2.11	**1.61**
(2014-02-28 - 2019-01-24)	h=5	2.32	2.07	**1.61**
	h=20	2.05	2.33	**1.62**
Subsample 1	h=1	2.23	1.94	**1.83**
(2014-02-28 - 2016-02-25)	h=5	2.18	2.24	**1.58**
	h=20	2.11	2.36	**1.53**
Subsample 2	h=1	2.06	2.33	**1.61**
(2016-02-26 - 2019-01-24)	h=5	2.38	1.95	**1.67**
	h=20	2.00	2.26	**1.74**

*Notes: See notes to Table 3. Entries in this table are average rankings for all forecasting models (AR+PCA, AR+SPCA, AR+HT) at forecast horizons of 1, 5, and 20 days ahead, across the full sample and two subsamples, with "1" being the best model.

Table 5: Comparison of Forecasts in Model Combinations by Number of Stocks with Lowest MSFE Across Forecast Horizons and Subsamples*

		Comb 1	Comb 2	Best Single
Full Sample	h=1	40	**259**	92
(2014-02-28 - 2019-01-24)	h=5	49	**248**	94
	h=20	68	**266**	57
Subsample 1	h=1	53	**199**	139
(2014-02-28 - 2016-02-25)	h=5	55	**257**	79
	h=20	99	6	**286**
Subsample 2	h=1	67	**273**	51
(2016-02-26 - 2019-01-24)	h=5	124	51	**216**
	h=20	75	**249**	67

*Notes: See notes to Table 4. Entries in this table are the number of stocks for which each model combination or individual model (called Best Single) has the lowest MSFE among all models at forecast horizons of 1, 5, and 20 days ahead, across the full sample and two subsamples. The model combinations are detailed in Table 2.

Table 6: Table of Average Rankings of All Single Models Across Forecast Horizons and Subsamples*

		Comb1	Comb2	Best Single
Full Sample	h=1	2.28	**1.62**	2.10
(2014-02-28 - 2019-01-24)	h=5	2.35	**1.62**	2.03
	h=20	2.10	**1.57**	2.33
Subsample 1	h=1	2.18	**1.87**	1.95
(2014-02-28 - 2016-02-25)	h=5	2.21	**1.57**	2.22
	h=20	2.48	1.99	**1.53**
Subsample2	h=1	2.08	**1.57**	2.35
(2016-02-26 - 2019-01-24)	h=5	2.32	1.89	**1.79**
	h=20	2.03	**1.69**	2.28

*Notes: See notes to Table 5. Entries in this table are average rankings, as discussed in the footnote to Table 4.