ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting crude oil market volatility using variable selection and common factor

Yaojie Zhang [a], M.I.M. Wahab [b,*], Yudong Wang [a]

[a] School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China
[b] Department of Mechanical and Industrial Engineering, Ryerson University, Toronto, Canada

ARTICLE INFO

Keywords:
Volatility forecasting
Crude oil market
Machine learning
Big data
Variable selection

ABSTRACT

This paper aims to improve the predictability of aggregate oil market volatility with a substantially large macroeconomic database, including 127 macro variables. To this end, we use machine learning from both the variable selection (VS) and common factor (i.e., dimension reduction) perspectives. We first use the lasso, elastic net (ENet), and two conventional supervised learning approaches based on the significance level of predictors' regression coefficients and the incremental *R*-square to select useful predictors relevant to forecasting oil market volatility. We then rely on the principal component analysis (PCA) to extract a common factor from the selected predictors. Finally, we augment the autoregression (AR) benchmark model by including the supervised PCA common index. Our empirical results show that the supervised PCA regression model can successfully predict oil market volatility both in-sample and out-of-sample. Also, the recommended models can yield forecasting gains in both statistical and economic perspectives. We further shed light on the nature of VS over time. In particular, option-implied volatility is always the most powerful predictor.

1. Introduction

Undoubtedly, crude oil is a major commodity in the financial market due to its financialization. It is commonly recognized that the oil price uncertainty significantly influences the equity market and economic activities (see, e.g., Elder & Serletis, 2010; Ferderer, 1996; Hou et al., 2016; Jo, 2014; Kilian & Park, 2009; Kilian & Vigfusson, 2017; Wang et al., 2013). Also, access to the accurate forecast of oil market volatility is essential to portfolio allocation, risk management, and asset pricing.[1] Hence, modeling and predicting oil price volatility are of great interest to academics and practitioners. As a result, a substantially large number of articles investigate the predictability of oil market volatility (see, e.g., Degiannakis & Filis, 2017; Gong & Lin, 2018; Haugom et al., 2014; Kristjanpoller & Minutolo, 2016; Liang, Wei, Li, et al., 2020; Ma, Ji, et al., 2019; Ma, Liao, et al., 2019; Ma et al., 2018; Pan et al., 2017; Sévi, 2014; Wei et al., 2017; Zhang, Wei, et al., 2019). In particular, this paper contributes to the extant literature on oil market volatility prediction by shedding new light on the big data and machine learning contexts.

This study relies on a comprehensive macroeconomic database, the FRED-MD, established by McCracken and

* Correspondence to: Department of Mechanical and Industrial Engineering, Ryerson University, 350 Victoria Street, Toronto Ontario, M5B 2K3, Canada.

E-mail addresses: yaojie_zhang@126.com (Y. Zhang), wahab@ryerson.ca (M.I.M. Wahab), wangyudongnj@126.com (Y. Wang).

[1] Likewise, the accurate forecasts of other financial market volatility are equally important to these applications. Therefore, a voluminous literature also tries to improve the volatility forecast accuracy for a wide range of financial markets (see, e.g., Bollerslev et al., 2018; Li et al., 2019; Nonejad, 2017; Paye, 2012; Wang et al., 2020, 2018; Zhang et al., 2020).

https://doi.org/10.1016/j.ijforecast.2021.12.013
0169-2070/

Please cite this article as: Y. Zhang, M.I.M. Wahab and Y. Wang, Forecasting crude oil market volatility using variable selection and common factor. International Journal of Forecasting (2022), https://doi.org/10.1016/j.ijforecast.2021.12.013.

Ng (2016), facilitating "big-data"[2] research in macroeconomic forecasting. We obtain a total of 127 macro variables via the FRED-MD database. The conventional multiple regression models dealing with such a large set of predictors presumably suffer from overfitting, multicollinearity, and the curse of dimensionality. This paper follows Bai and Ng (2008) and depends on machine learning techniques for the variable selection (VS) and dimension reduction (i.e., common factor) to overcome the issues with conventional multiple regression models.

We first rely on four supervised learning methods to parsimoniously select predictors that are relevant to forecasting oil market volatility. Subsequently, we use the principal component analysis (PCA) to construct a common factor from the selected predictors rather than all potential predictors. The conventional PCA technique is a widely used machine learning approach. However, it ignores the target information (i.e., oil market volatility). Hence, the conventional PCA technique is a simple dimension reduction method without supervised learning. To achieve both supervised learning and dimension reduction, we combine the machine learning approaches of VS and common factor. Among the four supervised learning methods we use, the first two are traditional supervised learning methods. They are based on the t-statistic for the regression slope of a potential predictor, above a significance level (termed t-stat), and sizable incremental R^2 relative to the autoregression (AR) benchmark model (termed ΔR^2). The others are two popular shrinkage methods: lasso (Tibshirani, 1996) and elastic net (ENet) (Zou & Hastie, 2005).

Following the related literature on the prediction of aggregate market volatility (see, e.g., Christiansen et al., 2012; Nonejad, 2017; Paye, 2012; Wang et al., 2018), we build our empirical analysis on the AR framework, since aggregate market volatility is highly persistent. The AR benchmark model can capture most of its predictive components. Furthermore, we develop the supervised PCA regression models by separately adding the PCA common factors with supervised learning (i.e., VS) into the AR benchmark. To demonstrate the superiority of the supervised PCA regression models in forecasting oil market volatility, we consider the AR benchmark and some competing models. First, the lasso, ENet, and conventional PCA regression model based on all the predictors without VS are essential competing models. We also consider the so-called kitchen sink model that adds all the potential predictors into the AR model. More moderately, we further use a few refinements to the kitchen sink model incorporating the relevant predictors selected by t-stat, ΔR^2, lasso, or ENet to the AR model.

The empirical results suggest that the supervised PCA regression models with VS outperform the AR benchmark both in-sample and out-of-sample across 1-, 3-, 6- and 12-month horizons. We also find that the in-sample and out-of-sample predictability increases with the forecast horizon. However, the other competing models fail to beat the simple AR benchmark out-of-sample. Specifically, for the in-sample predictability, supervised PCA indexes (i.e., common factors) based on the selected predictors yield significant regression slopes and sizable incremental R^2 relative to the AR benchmark, ranging from 0.811% to 5.689%. The Campbell and Thompson (2008) out-of-sample R^2 (R^2_{OS}) statistic assesses the out-of-sample predictive performance. Only the supervised PCA regression models with VS yield significantly positive R^2_{OS}s, thus beating the benchmark and other competing models. Surprisingly, the R^2_{OS}s for our recommended models at the 1-, 3-, 6- and 12-month horizons are around 1.5%, 2%, 5%, and 6%, respectively. This is an impressive out-of-sample forecasting performance. Furthermore, the forecasting results are robust to alternative volatility measures based on the oil futures and spot markets, alternative settings for VS, alternative model specifications for the AR benchmark, and many other situations (see the Online Appendix).

Our forecasting results echo the findings of the M4 Competition (see Makridakis et al., 2018). There are five major findings for the M4 Competition. We review two of them, which are related to our findings. First, the M4 Competition suggests that a hybrid approach combining statistical and machine learning elements comes first among many competing models. This echoes our finding that our best strategy is a hybrid of VS and dimension reduction, involving statistical and machine learning elements. Second, the M4 Competition finds that six pure machine learning methods show relatively poor forecasting performance. This is consistent with our finding that the pure lasso, ENet, and PCA approaches cannot beat the simple AR benchmark. We believe that hybrid approaches combining different elements are notable for future research, given the two similar findings.

We also implement a portfolio exercise proposed by Bollerslev et al. (2018) to examine the economic value of the volatility forecasts generated by the recommended machine learning approaches. Compared with the benchmark and competing forecasts, the supervised PCA regression forecasts produce the highest utility for a mean–variance investor who allocates her portfolio between crude oil futures and risk-free bills. Hence, the recommended machine learning approaches can produce forecasting gains from both statistical and economic perspectives.

We further shed light on the nature of selected variables relevant to forecasting oil market volatility over time. We observe that VS changes over time. A relatively large number of predictors are recursively selected over the entire out-of-sample period. Specifically, the traditional supervised learning approaches of t-stat and ΔR^2 typically select variables from the same categories, showing a clustered selection pattern. In contrast, the lasso and ENet tend to select diverse predictors from different categories, showing a scattered selection pattern. This can be explained by the statistical inference of the lasso and ENet because the use of the L_1 penalty will result in a strong tendency to select one predictor from many

[2] Goldstein et al. (2021) argue that the definition of big data in finance research should be different from the one in engineering and statistics. They propose three properties for big data in finance: large size, high dimension, and complex structure. This paper fits into the category of high dimension, which involves lots of variables. In this sense, we label the FRED-MD database as "big-data", as in Huang et al. (2021).

similar predictors. Overall, all four supervised learning approaches show an overwhelming tendency to select the variables from the stock market category, especially the stock market volatility index (i.e., an option-implied volatility variable). This echoes the closely associated volatility behaviors between the crude oil and equity markets, as well as oil financialization (see, e.g., Arouri et al., 2011, 2012; Creti et al., 2013; Degiannakis & Filis, 2017; Wang et al., 2018; Ma, Ji, et al., 2019). Option-implied volatility is the most powerful predictor, with a fairly high selection frequency at 98.4% over the out-of-sample forecasting period. This finding is supported by a large body of literature on volatility forecasting (see, e.g., Busch et al., 2011; Jeon et al., 2020; Liang, Wei, & Zhang, 2020; Seo & Kim, 2015). This is probably because option-implied volatility contains future information from informed traders.

This paper is related to the literature on macroeconomic variables and aggregate market volatility forecasting (Christiansen et al., 2012; Nonejad, 2017; Paye, 2012). We contribute to the related literature from at least three aspects. First, we extend their studies on the stock market to an equally important case, namely the crude oil market. Second, our empirical analysis is based on a substantially more extensive set of macro variables, which helps investigate the role of big data in financial forecasting. Third, and perhaps more importantly, Paye (2012) documents that individual macro variables cannot significantly forecast aggregate equity market volatility out-of-sample. To improve the forecasting performance, Paye (2012) relies on simple forecast combinations, while Christiansen et al. (2012) and Nonejad (2017) use a sophisticated combination approach of the Bayesian model averaging. Unlike theirs, we use the machine learning approaches from the VS and common factor perspectives and provide new insights.

The recent papers by Liang, Wei, Li, et al. (2020) and Ma et al. (2018) perhaps come closest to our work. Liang, Wei, Li, et al. (2020) and Ma et al. (2018) rely on the lasso to improve the forecasting performance for aggregate oil market volatility. However, they depend on a relatively small set of predictors. In contrast, our study is built on a "big-data" environment. Furthermore, we find that it is no longer feasible to depend merely on the lasso in our data-rich context. We complement their studies by combining the machine learning approaches with VS and common factor. We provide evidence that our methods can successfully address a plethora of plausible variables and thus help to enhance the predictability of aggregate oil market volatility.

More broadly, our work is related to machine learning and financial forecasting literature. Gu et al. (2020) use dimension reduction methods, boosted regression trees, random forests, penalized regressions, and neural networks to forecast equity risk premiums. Zhang and Hamori (2020) rely on random forests, logistic regression, support vector machines, and extreme gradient boosting (XGBoost) to predict crude oil price crashes. We complement their work by providing a hybrid perspective of machine learning. Furthermore, we provide valuable insights regarding VS into the specific forecasting field of crude oil market volatility. This contributes to understanding the behavior of crude oil price volatility.

The rest of the paper is outlined as follows. Section 2 provides the forecasting methodology. Section 3 details the data. Section 4 presents the in-sample and out-of-sample results and some robustness checks. Section 5 investigates the economic significance of volatility forecasts from a portfolio allocation perspective. Section 6 sheds light on the nature of VS. Section 7 concludes.

2. Methodology

2.1. Realized variance and benchmark model

Following the literature on aggregate market volatility forecasting (see, e.g., Christiansen et al., 2012; Nonejad, 2017; Paye, 2012; Wang et al., 2018), we calculate the realized variance (RV) as a proxy for the aggregate oil market volatility. Specifically, RV is calculated as the summation of daily squared returns,

$$RV_t = \sum_{j=1}^{M_t} r_{t,j}^2, \tag{1}$$

where RV_t denotes the realized variance for month t, M_t denotes the number of all trading days in month t, and $r_{t,j}$ represents the daily return for the jth trading day in month t.

The aggregate market volatility calculated by Eq. (1) is typically leptokurtotic and non-Gaussian. Most of the predictive regressions used by this paper are estimated by ordinary least squares (OLS). Nonetheless, the OLS estimator appears to be misleading when regression errors are non-Gaussian. To avoid such an issue, we take the natural logarithm of the original RV, i.e., $LV_t = \ln(RV_t)$, which is approximately Gaussian (see also Christiansen et al., 2012; Nonejad, 2017; Paye, 2012; Wang et al., 2018). Hence, in the empirical analysis, we rely on the logarithmic RV, LV_t, to measure and forecast the aggregate oil market volatility.

The prevailing benchmark for forecasting aggregate market volatility is the AR model,

$$LV_{t+1:t+h} = \alpha + \sum_{\ell=1}^{L(h)} \beta_\ell LV_{t+1-\ell} + \omega_{t+1:t+h}, \tag{2}$$

where $LV_{t+1:t+h} = \ln[(1/h)(RV_{t+1} + \cdots + RV_{t+h})]$ denotes the market variance from month t + 1 to month t + h, h represents the forecast horizon, and $\omega_{t+1:t+h}$ denotes the regression error. The optimal lag length of $L(h)$ is determined by the Akaike information criterion (AIC).[3] Our empirical sample suggests that $L(h) = 2$, 6, 6, and 5 when $h = 1$, 3, 6, and 12, respectively. Paye (2012) and Wang et al. (2018) also use the AR benchmark model.

[3] In the robustness check below, we further consider the Bayesian information criterion (BIC) and adjusted R^2 to choose the optimal number of AR lags. Our results are not specific to alternative choices of AR lags.

2.2. Variable selection

This study has N ($N = 127$) macroeconomic variables to forecast oil futures market volatility. The individual predictive regression based on each macroeconomic variable is expressed as follows:

$$LV_{t+1:t+h} = \alpha + \sum_{\ell=1}^{L(h)} \beta_\ell LV_{t+1-\ell} + \varphi_i X_{i,t} + \omega_{t+1:t+h}, \quad \text{for } i = 1, 2, \ldots, N, \tag{3}$$

where $X_{i,t}$ is the ith predictor. This is an augmented AR model.

Bai and Ng (2008) propose hard and soft thresholding approaches to select targeted variables. The hard thresholding approach selects powerful predictors based on the t-statistics (t-stat) for their regression slopes φ_i in Eq. (3). To obtain a moderate size for the selected predictor subset, we depend on the threshold value of t-stat at the 10% two-tailed level, and it is around 1.65.[4] Analogously, we additionally rely on the predictive R^2 to select the useful predictors. More specifically, we separately run the AR benchmark model in Eq. (2) and the augmented AR model with the ith macroeconomic predictor in Eq. (3). Then, the incremental R^2, ΔR^2, is calculated as the R^2 of regression (3) minus the R^2 of regression (2). We select the ith predictor if the ΔR^2 is above the threshold value of 1%. Hence, the selected predictor can provide complementary forecasting information beyond the one already contained in the lags of the dependent variable (i.e., aggregate oil market volatility). Such information content is thus helpful in improving the predictability of oil futures market volatility. Finally, it is noteworthy that we select no less than 5 predictors to construct the PCA index. If no more than 5 predictors satisfy the cutoff point for t-stat or ΔR^2, we just select 5 predictors with the largest five values of t-stat or ΔR^2. According to the evidence of VS shown below, this reasonable choice hardly influences our forecasting results.

A defect of hard thresholding is that this method only concentrates on a specific predictor but ignores the associated information with other predictors. In contrast, the soft thresholding approach overcomes this issue. Specifically, we use two prevailing shrinkage methods of lasso (Tibshirani, 1996) and ENet (Zou & Hastie, 2005). The lasso and ENet perform VS and regularization to enhance the interpretability and predictability of regression models from a machine learning perspective. More importantly, when a few influential variables deliver overlapping information, the lasso and ENet tend to select the most representative one and discard the remaining ones.

Statistically, the lasso and ENet are based on the following regression model:

$$LV_{t+1:t+h} = \alpha + \sum_{\ell=1}^{L(h)} \beta_\ell LV_{t+1-\ell} + \sum_{i=1}^{N} \varphi_i X_{i,t} + \omega_{t+1:t+h}. \tag{4}$$

Also, this is the kitchen sink (KS) model. The KS forecast is generated by the OLS estimates. However, the lasso forecast is produced by the lasso coefficient estimates based on the following objective function:

$$\underset{\alpha,\beta_1,\ldots,\beta_{L(h)},\varphi_1,\ldots,\varphi_N \in \mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{2(t-L(h)+1)} \sum_{\tau=L(h)}^{t} \left[LV_{\tau+1:\tau+h} - \left(\alpha + \sum_{\ell=1}^{L(h)} \beta_\ell LV_{\tau+1-\ell} + \sum_{i=1}^{N} \varphi_i X_{i,\tau} \right) \right]^2 + \lambda \left(\sum_{\ell=1}^{L(h)} |\beta_\ell| + \sum_{i=1}^{N} |\varphi_i| \right) \right\}, \tag{5}$$

where λ is the regularization parameter, which is nonnegative and controls the degree of shrinkage. The lasso uses the L_1 parameter penalization so that it can set regression coefficients to exactly zero and thereby performs VS. In contrast, the ridge regression, pioneered by Hoerl and Kennard (1970), imposes the L_2 parameter penalization on regression coefficients. It thus only shrinks the magnitude of coefficients (but does not shrink them to exactly zero). Based on both the L_1 and L_2 penalties, the ENet shrinks regression coefficients using the following objective function:

$$\underset{\alpha,\beta_1,\ldots,\beta_{L(h)},\varphi_1,\ldots,\varphi_N \in \mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{2(t-L(h)+1)} \sum_{\tau=L(h)}^{t} \left[LV_{\tau+1:\tau+h} - \left(\alpha + \sum_{\ell=1}^{L(h)} \beta_\ell LV_{\tau+1-\ell} + \sum_{i=1}^{N} \varphi_i X_{i,\tau} \right) \right]^2 + \lambda \left[\rho \left(\sum_{\ell=1}^{L(h)} |\beta_\ell| + \sum_{i=1}^{N} |\varphi_i| \right) + 0.5\,(1-\rho) \left(\sum_{\ell=1}^{L(h)} \beta_\ell^2 + \sum_{i=1}^{N} \varphi_i^2 \right) \right] \right\}, \tag{6}$$

where ρ is a parameter in the range of 0 and 1, which balances the L_1 and L_2 penalties. For simplicity, we follow Rapach and Zhou (2020) and set ρ to 0.5.[5] In terms of choice for λ, we rely on the corrected version of AIC (Hurvich & Tsai, 1989). Flynn et al. (2013) provide evidence that the corrected AIC has good asymptotic and finite-sample properties to determine the desired value of λ.[6]

4 In the out-of-sample forecasting procedure, the threshold value of t-stat changes slightly with an increase in the number of in-sample estimation observations since the degree of freedom increases accordingly.

5 The results are qualitatively similar when we use the corrected AIC to choose the value of ρ. To reduce the huge computational burden, we simply set ρ to 0.5.

6 The forecasting results are qualitatively similar for alternative methods to determine the parameters of λ and ρ. For more details, see the corresponding robustness check below.

2.3. Principal component analysis and its supervised version

To alleviate the concern of overfitting, the PCA is a widely used machine learning technique for the time-series predictability (see, e.g., Degiannakis & Filis, 2017; Ludvigson & Ng, 2007, 2009; McCracken & Ng, 2016; Neely et al., 2014; Zhang & Hamori, 2020). Also, the PCA can reduce the dimension and thus avoid the curse of dimensionality. Hence, the PCA is suitable for this study in a data-rich environment. Statistically, the conventional PCA constructs common factors f_t based on all the potential predictors of $(X_{1,t}, \ldots, X_{N,t})'$ by a linear factor structure,

$$X_{i,t} = \pi_i' f_t + \varepsilon_{i,t}, \tag{7}$$

where f_t denotes the K-dimensional latent common factors (i.e., principal components), π_i denotes the latent factor loadings, and $\varepsilon_{i,t}$ denotes the idiosyncratic error term. We can realize the role of considerable dimension reduction when $K \ll N$. Using the most important and powerful common factor, namely the first principal component F_t ($F_t \subset f_t$), we run the following PCA regression model:

$$LV_{t+1:t+h} = \alpha + \sum_{\ell=1}^{L(h)} \beta_\ell LV_{t+1-\ell} + \phi F_t + \omega_{t+1:t+h}. \tag{8}$$

The PCA index can not only filter out most idiosyncratic noises contained in individual predictor variables but also extract the most striking co-movement from all the potential variables. Nonetheless, the conventional PCA does not consider the target (i.e., the dependent variable). Hence, it is an unsupervised learning technique. In this paper, we rely on the technique of VS to achieve supervised learning for the PCA. The supervised PCA is more equipped to filter out both the idiosyncratic and common noise components. Specifically, the predictive regression model for the supervised PCA is the augmented AR model that further includes the common factor extracted from the selected variables (instead of all the potential variables), which takes the form of

$$LV_{t+1:t+h} = \alpha + \sum_{\ell=1}^{L(h)} \beta_\ell LV_{t+1-\ell} + \phi F_t^{\mathrm{VS}} + \omega_{t+1:t+h}$$
$$\text{for VS} = t\text{-stat},\ \Delta R^2,\ \text{lasso},\ \text{ENet}, \tag{9}$$

where F_t^{VS} is the first principal component extracted from the selected predictors using the VS techniques for VS $= t$-stat, ΔR^2, lasso, or ENet. More specifically, we have two steps before running the regression model in Eq. (9). In the first step, we use each VS technique to select useful predictors. In the second step, we construct the first principal component based on the selected predictors. The resulting F_t^{VS} is supposed to extract more useful forecasting information since the selected variables are more relevant to the target.

Finally, note that the supervised PCA regression models with VS in Eq. (9) are termed PCA-VS. More specifically, PCA-VS becomes PCA-t-stat, PCA-ΔR^2, PCA-lasso, or PCA-ENet when VS $= t$-stat, ΔR^2, lasso, or ENet, respectively.

2.4. Competing models

To demonstrate the superiority of the predictive ability for our recommended models (i.e., the supervised PCA with VS), we should consider some related competing models. Of course, the AR benchmark model is the most necessary competing model. In addition, the conventional PCA regression model in Eq. (8), which is dubbed PCA-all as it uses all the predictors, is another essential competing model. Likewise, the VS methods of lasso and ENet are equally essential competing models.

The last strand of competing models is the traditional predictive regression model with multiple predictors. The first example we consider is the KS model with all the potential predictors, as shown in Eq. (4). This model suffers from overfitting and multicollinearity problems when using a large set of predictors. In contrast, a possibly more feasible strategy is that we should moderately include the selected predictors into the multiple regression models, which take the form of

$$LV_{t+1:t+h} = \alpha + \sum_{\ell=1}^{L(h)} \beta_\ell LV_{t+1-\ell} + \sum_{i \in \mathbb{N}^{\mathrm{VS}}} \varphi_i X_{i,t} + \omega_{t+1:t+h}, \tag{10}$$

where $\mathbb{N}^{\mathrm{VS}}$ is the index set of the selected variables for VS $= t$-stat, ΔR^2, lasso, or ENet. Accordingly, the predictive models in Eq. (10) are dubbed KS-VS, namely KS-t-stat, KS-ΔR^2, KS-lasso, and KS-ENet when VS $= t$-stat, ΔR^2, lasso, and ENet, respectively. To differentiate the KS models further, the KS model that includes all the predictors given in Eq. (4) is called KS-all.

In summary, our recommended models are PCA-t-stat, PCA-ΔR^2, PCA-lasso, and PCA-ENet, while the competing models include lasso, ENet, PCA-all, KS-all, KS-t-stat, KS-ΔR^2, KS-lasso, and KS-ENet.[7] Table 1 summarizes all the forecasting models used in this paper.

2.5. Forecast evaluation

Following the convention in monthly prediction (see, e.g., Christiansen et al., 2012; Jiang et al., 2019; Neely et al., 2014; Rapach et al., 2016, 2010), we generate out-of-sample forecasts by employing an expanding estimation window.[8] In other words, VS, common factor, and

[7] The above-mentioned competing models are most related to our recommended models. We further consider other competing models, including time-varying parameter autoregression (TVP-AR) model, dynamic model averaging (DMA), dynamic model selection (DMS), Bayesian model averaging (BMA), Bayesian model selection (BMS), and Bayesian semiparametric selection (BSS) (see Koop & Korobilis, 2011, 2012, 2013; Korobilis, 2013a, 2013b; Raftery et al., 2010). The Online Appendix reports their forecasting results, which suggest that our recommended models continue to show significantly better forecasting performance than these competing models.

[8] The rolling estimation window is also widely used to generate volatility forecasts. The rolling estimation window is more popular for forecasting daily volatility, while the expanding window is more popular for forecasting monthly volatility. This is because the time series of daily volatility is vulnerable to structural breaks and the

Table 1
Summary of all the forecasting models.

Model name	Description	Equation
AR	the benchmark, an autoregression model with past oil market volatility variables	Eq. (2)
PCA-t-stat	the AR benchmark + the PCA common factor from the predictors selected by t-stat.	Eq. (9)
PCA-ΔR^2	the AR benchmark + the PCA common factor from the predictors selected by ΔR^2	Eq. (9)
PCA-lasso	the AR benchmark + the PCA common factor from the predictors selected by lasso	Eq. (9)
PCA-ENet	the AR benchmark + the PCA common factor from the predictors selected by ENet	Eq. (9)
Lasso	a penalized regression with the L_1 penalty	Eq. (5)
ENet	a penalized regression with both the L_1 and L_2 penalties	Eq. (6)
PCA-all	the AR benchmark + the PCA common factor from all the predictors	Eq. (8)
KS-all	the AR benchmark + all the predictors	Eq. (4)
KS-t-stat	the AR benchmark + the selected predictors via t-stat	Eq. (10)
KS-ΔR^2	the AR benchmark + the selected predictors via ΔR^2	Eq. (10)
KS-lasso	the AR benchmark + the selected predictors via lasso	Eq. (10)
KS-ENet	the AR benchmark + the selected predictors via ENet	Eq. (10)

model estimation are performed recursively. It is noteworthy that we only use variable information available up to month t when forecasting $LV_{t+1:t+h}$. By doing so, we can avoid the look-ahead bias in our out-of-sample predictions.

To evaluate forecast accuracy, we rely on the out-of-sample R^2 (R^2_{OS}) statistic, as in Campbell and Thompson (2008) and Wang et al. (2018). The R^2_{OS} measures the proportional reduction in the mean squared predictive error (MSPE) of a forecasting model of interest relative to that of the AR model. Statistically, the R^2_{OS} statistic is computed by the following expression:

$$R^2_{OS} = 1 - \frac{\sum_{t=p}^{T-h}\left(LV_{t+1:t+h} - \widehat{LV}^{M}_{t+1:t+h}\right)^2}{\sum_{t=p}^{T-h}\left(LV_{t+1:t+h} - \widehat{LV}^{B}_{t+1:t+h}\right)^2}, \tag{11}$$

where p is the number of observations in the initial training sample, T is the number of observations over the whole sample, $LV_{t+1:t+h}$ is the true value of the logarithmic RV for the period of $t+1$: $t+h$, and $\widehat{LV}^{M}_{t+1:t+h}$, and $\widehat{LV}^{B}_{t+1:t+h}$ are the forecasts of a forecasting model of interest and the AR benchmark model, respectively. A positive R^2_{OS} implies that the corresponding predictive model yields more accurate forecasts than the AR benchmark model, while a negative value indicates the opposite. Furthermore, the statistical significance for R^2_{OS} is tested by the Clark and West (2007) statistic, which is termed CW hereafter. The CW statistic tests the null hypothesis that the MSPE of the AR benchmark is lower than or equal to the MSPE of a predictive model against the upper-tail alternative hypothesis that the MSPE of the AR benchmark is higher than the MSPE of a predictive model.[9]

rolling window can mitigate this issue. However, the convention in monthly prediction is to use the expanding window because a monthly time series is less vulnerable to structural breaks. Furthermore, compared to the rolling window, the expanding window can use more observations to precisely estimate the initial parameters. This feature is suitable for monthly prediction as monthly variables always have limited observations. The Online Appendix reports the comparison results, which suggest that the expanding window is a better choice.

3. Data

In this paper, we rely on a large number of macroeconomic variables from the FRED-MD database of McCracken and Ng (2016) to forecast the aggregate oil futures market volatility. The FRED-MD is a comprehensive macroeconomic database consisting of more than 100 macro variables and facilitates "big-data" research. We collect a total of 128 macro variables from the 2019:06 vintage of the FRED-MD database.[10] The variable of ACOGNO, which is new orders for consumer goods, is excluded as this variable is merely available starting from February 1992. Consequently, we rely on 127 variables to forecast oil market volatility. The macro variables in the FRED-MD are classified into eight groups: (1) output and income (No. 1–16), (2) labor market (No. 17–47), (3) housing (No. 48–57), (4) consumption, orders, and inventories (No. 58–66), (5) money and credit (No. 67–80), (6) interest and exchange rates (No. 81–102), (7) prices (No. 103–122), and (8) stock market (No. 123–127). Two additional issues are noteworthy. First, we use the transformation codes provided by McCracken and Ng

9 The forecasting results are similar when we use the Diebold and Mariano (1995) test based on the loss functions of QLIKE, HMSE, and HMAE (see the Online Appendix).

10 The data for the FRED-MD are available from Dr. Michael McCracken's homepage at https://research.stlouisfed.org/econ/mccracken/sel/. See the FRED-MD Appendix of McCracken and Ng (2016) for an informative list involving variable descriptions and transformation codes.

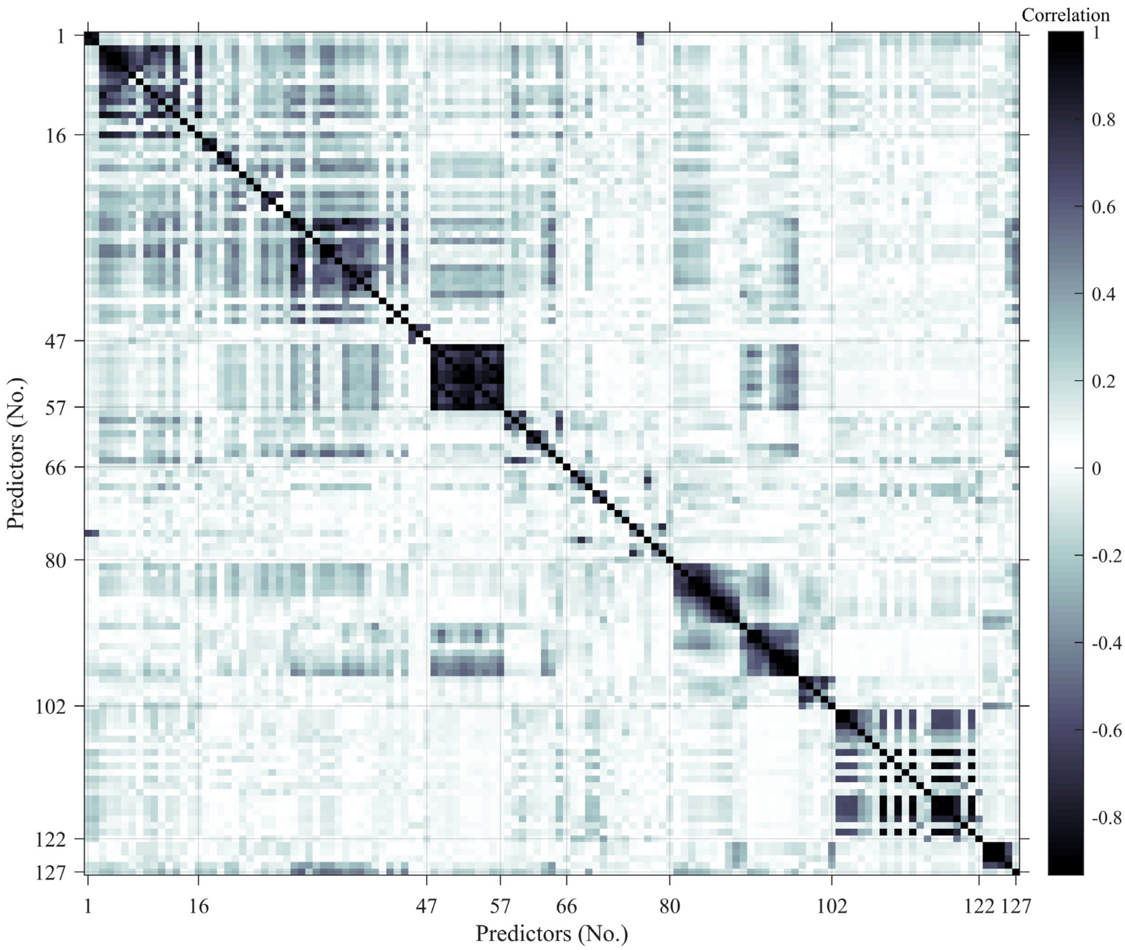

Fig. 1. Correlations for all the 127 used predictors.
This figure depicts a heatmap for the correlations among the 127 FRED-MD predictors. All the predictors are classified into eight groups: (1) output and income (No. 1–16), (2) labor market (No. 17–47), (3) housing (No. 48–57), (4) consumption, orders, and inventories (No. 58–66), (5) money and credit (No. 67–80), (6) interest and exchange rates (No. 81–102), (7) prices (No. 103–122), and (8) stock market (No. 123–127).

(2016) to ensure the stationarity for each used variable. Second, we take a one-month lag for many variables due to the publication delays.

The aggregate oil futures market volatility is calculated based on daily market returns. We collect the daily price data for the West Texas Intermediate (WTI) futures from the U.S. Energy Information Administration (EIA).[11] WTI is a well-known oil price benchmark. While international investors pay extensive attention to WTI, the WTI price is more susceptible to the US domestic oil supply, demand, and many other macroeconomic fundamentals of the US than the other popular oil price benchmark of Brent. Given this, WTI is more suitable for the FRED-MD database. Our whole sample period runs from January 1985 to December 2018. The observations during the first 13 years are used as the initial training sample. Accordingly, the out-of-sample forecasting period spans January 1998 to December 2018.[12]

Fig. 1 depicts a heatmap for the correlations among the 127 FRED-MD predictors.[13] In the heatmap, a darker color implies a higher correlation for both positive and negative perspectives. A striking feature follows the figure immediately. The predictors in the same group tend to show strong correlations. In contrast, the ones in different groups are more likely to show weak correlations. This evidence suggests that different variable groups probably contain complementary information.

11 The data for daily oil prices are available from the webpage of the EIA at https://www.eia.gov/.

12 We further consider different window sizes to verify the robustness of our forecasting results (see the Online Appendix).

13 To save space, we report the summary statistics of the 127 FRED-MD variables and oil RV in the Online Appendix. Also, the Online Appendix plots the time series of oil RV.

Table 2
In-sample estimation results.

Models	$h = 1$		$h = 3$		$h = 6$		$h = 12$	
	ϕ/t-stat	ΔR^2 (%)	ϕ/t-stat	ΔR^2 (%)	ϕ/t-stat	ΔR^2 (%)	ϕ/t-stat	ΔR^2 (%)
PCA-t-stat	1.102	0.812	1.175	1.219	1.511	2.505	1.872	5.471
	(2.848)		(2.834)		(2.856)		(3.245)	
PCA-ΔR^2	1.102	0.812	1.176	1.220	1.506	2.463	1.873	5.471
	(2.847)		(2.834)		(2.819)		(3.255)	
PCA-lasso	1.101	0.811	1.175	1.217	1.520	2.526	1.916	5.689
	(2.846)		(2.832)		(2.862)		(3.256)	
PCA-ENet	1.101	0.811	1.175	1.217	1.506	2.463	1.916	5.688
	(2.846)		(2.832)		(2.819)		(3.256)	
PCA-all	−0.002	0.001	−0.019	0.255	−0.019	0.304	−0.013	0.194
	(−0.119)		(−2.018)		(−1.849)		(−1.104)	

This table presents the in-sample estimation results for the supervised (or unsupervised) PCA regression models, which take the form of

$$LV_{t+1:t+h} = \alpha + \sum_{\ell=1}^{L(h)} \beta_\ell LV_{t+1-\ell} + \phi F_t^{\mathrm{VS}}(\text{or } F_t) + \omega_{t+1:t+h},$$

where h denotes the forecast horizon, LV_t is the logarithmic realized variance for month t, F_t is the common factor extracted from all the predictors, and F_t^{VS} is the common factor extracted from the selected predictors using the variable selection (VS) techniques for VS = t-stat, ΔR^2, lasso, or ENet. We report the regression coefficients (ϕ) of common factors, Newey–West t-statistics (t-stat, in parentheses below), and the incremental R^2 (ΔR^2) relative to that of the AR benchmark. The entire sample period is 1985:01–2018:12.

4. Empirical results

4.1. In-sample estimation results

Table 2 reports the full-sample estimation results for the conventional PCA regression model (i.e., PCA-all) and the supervised PCA regression models with VS (i.e., PCA-VS). For the sake of brevity, we only report the regression coefficients of the PCA indexes (ignoring the AR lags), their Newey–West t-statistics, and the incremental R^2 (ΔR^2) relative to that of the AR benchmark. The regression coefficients of the supervised PCA indexes (i.e., the PCA-VS common factors) are all significant at the 1% level. In contrast, the coefficients of the conventional PCA index (i.e., the PCA-all common factors) show lower significance or even no significance. The ΔR^2s of PCA-VS vary from 0.811% to 5.689% with the increase of the forecast horizon. The ΔR^2s are sizable because the dependent variable of market volatility is highly persistent, and most of its future components have been explained by the lags in the AR benchmark. We thus conclude that PCA-VS common factors can provide considerable complementary information. Nonetheless, the conventional PCA-all model delivers relatively small ΔR^2s, thus failing to provide useful complementary information. Overall, the in-sample results suggest that the supervised PCA regression models with VS show powerful in-sample predictability for oil futures market volatility.

The results of the in-sample predictability imply that many macroeconomic variables are informative regarding future oil price volatility. This is intuitive since economic development would increase crude oil demand (Hamilton, 2009). In particular, Kilian (2009) presents that macroeconomic conditions always cause crude oil price changes due to crude oil demand changes. Kilian and Hicks (2013) also argue that unexpected economic growth is likely to raise oil prices. The theoretical basis and economic intuition lay a foundation for the success of macroeconomic variables for predicting oil price volatility.

4.2. Out-of-sample forecasting performance

The good in-sample predictability is possibly caused by overfitting. In contrast, the out-of-sample predictability is a relatively stringent test. The empirical literature on forecasting aggregate market return and volatility documents that a few individual predictive models can surpass the benchmark model in-sample, but they cannot continue to surpass the benchmark model out-of-sample (see, e.g., Paye, 2012; Welch & Goyal, 2008). Besides, it should be noted that the out-of-sample forecast accuracy is of more interest to investors in the practice of trading securities. In light of these crucial facts, we investigate out-of-sample forecasting performance in the following analysis.

Table 3 reports the out-of-sample forecasting results for all of the considered models. Three major findings emerge. First, only the four PCA-VS models (i.e., PCA-t-stat, PCA-ΔR^2, PCA-lasso, and PCA-ENet) yield significantly positive R^2_{OS}s, suggesting that the recommended models outperform the AR benchmark model. Second, the longer the forecast horizon, the greater R^2_{OS} for the PCA-VS models. Hence, the forecasting gains of the PCA-VS models increase with the forecast horizon. Specifically, the R^2_{OS}s for the PCA-VS models at the 1-, 3-, 6- and 12-month horizons are around 1.5%, 2%, 5%, and 6%, respectively, indicating very strong predictive power. Third, all the competing models fail to beat the AR benchmark model. In particular, the KS-all model yields the worst out-of-sample performance with extremely negative R^2_{OS}s, which are smaller than −250%. This is intuitive since the KS-all model severely suffers from overfitting and multicollinearity problems since all 127 predictors are included. The KS-VS models (i.e., KS-t-stat, KS-ΔR^2, KS-lasso, and KS-ENet) only incorporate the selected predictors and thus enhance the forecasting performance relative to the KS-all model. However, the KS-VS models still suffer from overfitting and cannot outperform the AR

Table 3
Out-of-sample forecasting performance.

Models	$h = 1$		$h = 3$		$h = 6$		$h = 12$	
	R^2_{OS} (%)	CW-stat	R^2_{OS} (%)	CW-stat	R^2_{OS} (%)	CW-stat	R^2_{OS} (%)	CW-stat
PCA-t-stat	2.02**	2.23	1.84***	2.73	4.57***	3.97	6.72***	6.02
PCA-ΔR^2	1.05**	1.91	1.93***	2.72	5.19***	4.06	6.74***	6.02
PCA-lasso	1.68**	2.14	3.09***	3.00	4.79***	4.00	5.62***	5.17
PCA-ENet	1.77**	2.17	2.51***	2.84	4.27***	3.87	8.12***	5.49
Lasso	−2.51	1.16	−8.08**	1.92	−28.58	0.05	−75.13	−1.56
ENet	−0.22**	1.65	−5.72**	2.25	−27.65	0.00	−63.78	−1.21
PCA-all	−1.11	−2.07	0.00	0.68	0.04	0.77	−0.41	0.12
KS-all	−252.42**	1.77	−4749.60	0.57	−42113.12	−0.64	−4451.41	−0.54
KS-t-stat	−15.40	0.91	−63.70	0.24	−109.83	−1.40	−127.37	−2.62
KS-ΔR^2	−12.78	1.10	−48.87	−0.18	−97.19	−1.00	−131.12	−2.74
KS-lasso	−17.99*	1.37	−38.50**	1.65	−58.76	0.03	−194.17	−2.16
KS-ENet	−14.21**	1.73	−31.76**	2.01	−60.69	−0.38	−196.31	−2.50

This table presents the out-of-sample R^2 (R^2_{OS}) for all the models relative to the AR benchmark. h denotes the forecast horizon. A positive R^2_{OS} implies that the corresponding predictive model outperforms the AR benchmark, while a negative value indicates the opposite. The statistical significance for R^2_{OS} is tested by the Clark and West (2007) statistic (CW-stat). *, **, and *** correspond to significance at the 10%, 5%, and 1% levels, respectively. The initial in-sample period is 1985:01-1997:12 and the out-of-sample evaluation period is 1998:01-2018:12.

benchmark model. Also, the machine learning approaches, such as the lasso, ENet, and PCA-all, yield larger R^2_{OS}s than those of the other competing models. However, only the PCA-all model at the 6-month horizon yields a small positive R^2_{OS}, merely 0.04%, with no significance. In summary, we can conclude that the recommended PCA-VS models show substantially better out-of-sample forecasting performance than the AR benchmark model and the related competing models.

4.3. Robustness checks

In this section, we consider some robustness checks to confirm the robustness of our out-of-sample forecasting results. First, we rely on the same forecasting models and predictors to forecast the WTI crude oil spot price volatility instead of its futures price volatility. Second, we consider alternative settings for VS methods. Third, we employ different criteria to choose the lags used by the AR benchmark model. Finally, the Online Appendix provides other robustness checks, such as alternative forecast evaluations, different window sizes, and other competing models.

4.3.1. Spot market volatility

Our empirical analysis focuses on the predictability of WTI futures market volatility. This subsection further investigates the out-of-sample predictive performance for the WTI spot market volatility. Daily WTI crude oil spot prices are also collected from the EIA. Since the spot price data started in January 1986, the whole sample period runs from January 1986 to December 2018. This is consistent with the prior analysis for the out-of-sample period. The same forecasting models and predictors are used to generate the forecasts of spot market volatility.

Table 4 reports the out-of-sample performance for forecasting the WTI spot market volatility. Consistent with the prior results in Table 3, only the PCA-VS models yield positive and significant R^2_{OS}s. A slight difference is that the forecasting gains of the supervised PCA regression models do not increase with the forecast horizon. The greatest R^2_{OS} value typically appears in the medium-term horizon. Nevertheless, our results are robust to futures or spot market volatility predictions.

4.3.2. Alternative settings for variable selection

The VS is central to this study. With this in mind, we further consider other possible ways to perform VS. The VS approaches using t-stat and ΔR^2 are based on the individual augmented AR models given in Eq. (3). Alternatively, we can regress the residual of the AR model in Eq. (2) on a constant and each lagged predictor. The value of t-stat is equal to the t-statistic of each predictor's regression coefficient, and the value of ΔR^2 is equal to the R^2 of the residual-based regression model.[14] Meantime, we employ the time-series-validation algorithm to determine the parameters of λ and ρ in the lasso and ENet. For more details about the time-series-validation algorithm, we refer the readers to Zhang, Ma, et al. (2019) and their Online Appendix.

Table 5 reports the out-of-sample forecasting results with alternative settings of VS. The results are comparable between Tables 3 and 5. The supervised PCA regression models with alternative VS methods continue to yield significantly positive R^2_{OS}s, thus beating the AR benchmark. However, all the competing models fail to beat the simple AR benchmark. Therefore, the forecasting results are not specific to the settings of VS. It is noteworthy that the use of four VS approaches (i.e., t-stat, ΔR^2, lasso, and ENet) also confirms the robustness of our findings for VS.

4.3.3. Alternative AR model specifications

In this paper, the AR model is a crucial benchmark to assess other models' predictive ability. Given this, we further consider alternative AR model specifications. In the previous analysis, we depend on the AIC to choose the number of lags of the dependent variables. However, the

[14] In terms of the threshold value for t-stat and ΔR^2, we also consider other reasonable values and obtain similar results. Due to space limitations, we omit the results but they are available upon request.

Table 4
Out-of-sample forecasting performance for spot market volatility.

Models	$h = 1$		$h = 3$		$h = 6$		$h = 12$	
	R^2_{OS} (%)	CW-stat	R^2_{OS} (%)	CW-stat	R^2_{OS} (%)	CW-stat	R^2_{OS} (%)	CW-stat
PCA-t-stat	1.99**	2.30	4.66***	3.13	4.52***	3.62	2.75***	5.11
PCA-ΔR^2	0.72**	1.85	3.85***	2.99	1.59***	3.12	2.74***	5.11
PCA-lasso	2.03**	2.30	3.40***	2.86	4.76***	3.66	2.56***	5.04
PCA-ENet	2.04**	2.32	3.68***	2.93	5.08***	3.73	2.68***	5.08
Lasso	−5.09	0.09	−4.26*	1.37	−20.06	−0.89	−75.34	−2.07
ENet	−5.32	−0.06	−5.47	1.05	−19.18	−0.87	−67.17	−1.86
PCA-all	−2.07	−1.17	−0.32	0.65	−0.34	0.32	−0.79	−0.37
KS-all	−560.57	0.66	−11039.04	0.27	−1273.94	−0.51	−8246.10	−0.01
KS-t-stat	−23.58	0.41	−30.48*	1.45	−78.42	−2.16	−162.22	−1.95
KS-ΔR^2	−11.15	0.38	−15.95**	2.28	−76.93	−1.68	−146.83	−1.73
KS-lasso	−19.52	1.18	−19.40*	1.46	−36.31	−0.38	−173.77	−1.97
KS-ENet	−20.32	0.72	−20.49	1.16	−38.49	−0.78	−119.10	−1.86

This table presents the out-of-sample R^2 (R^2_{OS}) for all the used models relative to the AR benchmark. h denotes the forecast horizon. In particular, we forecast the WTI spot market (instead of futures market) volatility in this table. A positive R^2_{OS} implies that the corresponding predictive model outperforms the AR benchmark, while a negative value indicates the opposite. The statistical significance for R^2_{OS} is tested by the Clark and West (2007) statistic (CW-stat). *, **, and *** correspond to significance at the 10%, 5%, and 1% levels, respectively. The initial in-sample period is 1986:01-1997:12 and the out-of-sample evaluation period is 1998:01-2018:12.

Table 5
Out-of-sample forecasting performance with alternative settings of variable selection.

Models	$h = 1$		$h = 3$		$h = 6$		$h = 12$	
	R^2_{OS} (%)	CW-stat	R^2_{OS} (%)	CW-stat	R^2_{OS} (%)	CW-stat	R^2_{OS} (%)	CW-stat
PCA-t-stat	1.50**	2.05	1.72***	2.70	4.53***	3.96	6.72***	6.02
PCA-ΔR^2	1.75**	2.17	3.15***	3.04	5.72***	4.22	6.72***	6.02
PCA-lasso	1.61**	2.11	3.58***	3.17	3.41***	3.67	7.38***	6.08
PCA-ENet	1.70**	2.15	3.80***	3.21	2.24***	3.34	5.05***	5.28
Lasso	−0.21*	1.53	−7.53	0.72	−25.54	−0.31	−61.55	−1.28
ENet	−0.35*	1.56	−8.49	0.71	−24.23	−0.16	−40.31	−0.19
PCA-all	−1.11	−2.07	0.00	0.68	0.04	0.77	−0.41	0.12
KS-all	−252.42**	1.77	−4749.60	0.57	−42113.12	−0.64	−4451.41	−0.54
KS-t-stat	−14.86	0.92	−65.58	0.06	−106.41	−1.30	−126.69	−2.66
KS-ΔR^2	−20.59	0.84	−70.30	0.55	−117.01	−1.45	−126.19	−2.67
KS-lasso	−18.58	0.79	−28.79	0.63	−50.70	−0.35	−84.51	−0.26
KS-ENet	−13.64*	1.34	−30.49	0.62	−59.28	−1.36	−95.85	−0.96

This table presents the out-of-sample R^2 (R^2_{OS}) for all the used models relative to the AR benchmark. h denotes the forecast horizon. In this table, we regress the residual of the AR model in Eq. (2) on a constant and each lagged predictor to obtain t-stat and ΔR^2; the lasso and ENet are performed by a time-series-validation algorithm. A positive R^2_{OS} implies that the corresponding predictive model outperforms the AR benchmark, while a negative value indicates the opposite. The statistical significance for R^2_{OS} is tested by the Clark and West (2007) statistic (CW-stat). *, **, and *** correspond to significance at the 10%, 5%, and 1% levels, respectively. The initial in-sample period is 1985:01-1997:12 and the out-of-sample evaluation period is 1998:01-2018:12.

Bayesian information criterion (BIC) and adjusted R^2 are also commonly adopted to choose the number of variables (see, e.g., Christiansen et al., 2012; Neely et al., 2014; Paye, 2012). For this consideration, we further employ the BIC and adjusted R^2 to choose the number of AR lags for different forecast horizons, i.e., $L(h)$. Given the $L(h)$ values obtained from the AIC, BIC, and adjusted R^2 together, we further set $L(h) = 3$, 2, 1, and 1 when $h = 1$, 3, 6, and 12, respectively.[15]

Table 6 reports the out-of-sample forecasting results with alternative AR model specifications. Yet again, the PCA-VS models exhibit substantially stronger predictive power than the AR benchmark and competing models. Therefore, our results are not specific to the AR model specifications.

5. Asset allocation

Financial practitioners may wonder whether the statistical forecasting gains of our models can accordingly produce economic gains. Hence, we further examine the economic value of the RV forecasts generated by the predictive models in an asset allocation exercise.[16] Many approaches rely on the return and volatility forecasts (see, e.g., Campbell & Thompson, 2008; Fleming et al., 2001, 2003; Rapach et al., 2010); (Zhang, Wei, et al., 2019). However, the portfolio exercise proposed by Bollerslev et al. (2018) exclusively depends on forecasting volatility. This is appealing since forecasting returns is notoriously difficult (see, e.g., Campbell & Thompson, 2008; Rapach

[15] The AIC evidence suggests that $L(h) = 2$, 6, 6, and 5 when $h = 1$, 3, 6, and 12, respectively; the BIC evidence suggests that $L(h) = 2$, 2, 1, and 1 when $h = 1$, 3, 6, and 12, respectively; the adjusted R^2 evidence suggests that $L(h) = 3$, 6, 6, and 6 when $h = 1$, 3, 6, and 12, respectively.

[16] In addition, we consider a hedging exercise. The results suggest that our recommended models show better hedging performance than the AR benchmark and all the competing models (see the Online Appendix). We are grateful to an anonymous referee for this interesting suggestion.

Table 6
Out-of-sample forecasting performance with alternative AR model specifications.

Models	$h = 1$		$h = 3$		$h = 6$		$h = 12$	
	R^2_{OS} (%)	CW-stat	R^2_{OS} (%)	CW-stat	R^2_{OS} (%)	CW-stat	R^2_{OS} (%)	CW-stat
PCA-t-stat	1.95***	2.37	0.98**	2.27	6.90***	4.04	5.87***	5.53
PCA-ΔR^2	1.94**	2.32	2.48***	2.59	4.29***	3.59	5.86***	5.53
PCA-lasso	1.48**	2.23	2.79***	2.79	3.09***	3.34	5.59***	5.41
PCA-ENet	1.76**	2.32	2.73***	2.66	8.00***	4.44	5.66***	5.43
Lasso	−2.53**	2.10	−6.76	0.96	−29.53	−1.15	−64.00	−2.20
ENet	−0.58***	2.48	−6.36	0.92	−28.20	−1.18	−62.81	−1.97
PCA-all	−1.11	−1.91	−0.33	0.23	0.01	0.75	−0.37	0.18
KS-all	−442.96**	1.80	−17668.41	−1.17	−150155.74	−0.68	−41230.23	−0.21
KS-t-stat	−19.17	0.83	−70.26	−1.33	−116.63	−1.81	−121.67	−2.70
KS-ΔR^2	−9.04**	1.81	−42.39	0.20	−114.73	−2.29	−128.57	−3.19
KS-lasso	−23.00*	1.62	−26.31	1.06	−48.54	−0.71	−105.15	−3.00
KS-ENet	−23.31*	1.40	−25.42	0.80	−56.68	−1.38	−117.34	−3.07

This table presents the out-of-sample R^2 (R^2_{OS}) for all the used models relative to the AR benchmark. h denotes the forecast horizon. In particular, we rely on the BIC and adjusted R^2 (rather than the AIC) to choose the number of AR lags in this table. A positive R^2_{OS} implies that the corresponding predictive model outperforms the AR benchmark, while a negative value indicates the opposite. The statistical significance for R^2_{OS} is tested by the Clark and West (2007) statistic (CW-stat). *, **, and *** correspond to significance at the 10%, 5%, and 1% levels, respectively. The initial in-sample period is 1985:01-1997:12 and the out-of-sample evaluation period is 1998:01-2018:12.

et al., 2010; Welch & Goyal, 2008). Accordingly, based on the work of Bollerslev et al. (2018), we assume that a mean–variance investor will allocate her portfolio between a risky asset (i.e., WTI crude oil futures) and a risk-free asset (i.e., risk-free bills) with a constant Sharpe ratio.

In the portfolio exercise of Bollerslev et al. (2018), an investor will invest a fraction w_t of her current (i.e., time t) portfolio in WTI oil futures with an excess return of r^e_{t+1} and the rest in risk-free bills with a risk-free return. The expected utility can be approximated as

$$U(w_t) = w_t E_t(r^e_{t+1}) - \frac{\gamma}{2} w_t^2 Var(r^e_{t+1}), \tag{12}$$

where γ denotes the investor's risk aversion coefficient and $Var(r^e_{t+1}) = E_t(RV_{t+1})$. To focus exclusively on volatility forecasting, Bollerslev et al. (2018) present the conditional Sharpe ratio, which is constant and can be expressed as $SR \equiv E_t(r^e_{t+1})/\sqrt{E_t(RV_{t+1})}$. Consequently, the expected utility can be rewritten as

$$U(w_t) = w_t SR\sqrt{E_t(RV_{t+1})} - \frac{\gamma}{2} w_t^2 E_t(RV_{t+1}), \tag{13}$$

which depends on the portfolio weight of w_t and the expected realized variance of $E_t(RV_{t+1})$. Maximizing the expected utility in Eq. (13), we can obtain the optimal portfolio weight for oil futures as follows.

$$w_t^* = \frac{SR/\gamma}{\sqrt{E_t(RV_{t+1})}}. \tag{14}$$

Given Eq. (14), we can derive that the conditional standard deviation of the portfolio's risky part is $\sqrt{Var(w_t^* r^e_{t+1})} = SR/\gamma$. This indicates that the investor targets the optimal volatility of SR/γ. When the forecast of $\sqrt{E_t(RV_{t+1})}$ is greater than the "risk target" of SR/γ (i.e., $w_t^* < 1$), the investor only allocates part of her wealth to the risky asset of oil futures. On the contrary, when the predicted volatility risk of $\sqrt{E_t(RV_{t+1})}$ is smaller than the risk target (i.e., $w_t^* > 1$), the investor has to rely on leverage to achieve her target.

Substituting Eq. (14) into Eq. (13), we can realize the expected utility regarding the optimally targeted portfolio as follows:

$$U(w_t^*) = \frac{SR^2}{2\gamma}. \tag{15}$$

However, in practice, $E_t(RV_{t+1})$ is not available. Using the RV forecast of $\widehat{RV}_{t+1}$ for month $t + 1$, we can realize the expected utility of

$$U(\widehat{RV}_{t+1}) = \frac{SR^2}{\gamma}\left(\frac{\sqrt{RV_{t+1}}}{\sqrt{\widehat{RV}_{t+1}}} - \frac{1}{2}\frac{RV_{t+1}}{\widehat{RV}_{t+1}}\right). \tag{16}$$

We empirically report the average utility during the entire out-of-sample forecasting period. Accordingly, the reported average utility is computed as

$$\overline{U}(\widehat{RV}) = \frac{1}{T-p}\sum_{t=p}^{T-1}\frac{SR^2}{\gamma}\left(\frac{\sqrt{RV_{t+1}}}{\sqrt{\widehat{RV}_{t+1}}} - \frac{1}{2}\frac{RV_{t+1}}{\widehat{RV}_{t+1}}\right). \tag{17}$$

Following Bollerslev et al. (2018), we set the risk aversion coefficient and annualized Sharpe ratio to be $\gamma = 2$ and $SR = 0.4$, respectively.[17] Consequently, $U(w_t^*) = 4\%$, implying that the investor is glad to pay 4% of her wealth to obtain the w_t^* risky asset portfolio rather than invest in risk-free bills exclusively. Similarly, we can obtain the average realized utility for longer forecast horizons.

Table 7 reports the portfolio performance evaluated based on the average realized utility. Consistent with the statistical evidence above, the economic results show that the PCA-VS models deliver greater utility values than the AR benchmark and the competing models. Most of the volatility forecasts result in positive utility, implying that using the volatility forecasts to allocate a portfolio is more profitable than simply investing in risk-free bills. More importantly, the PCA-t-stat, PCA-ΔR^2, PCA-lasso, and PCA-ENet consistently deliver the largest utility

[17] That is, the annualized volatility target equals 20%. Other reasonable values of SR and γ will not influence the comparison results of the average realized utility among different forecasting models.

Table 7
Portfolio performance.

Models	$h = 1$	$h = 3$	$h = 6$	$h = 12$
AR	3.547	3.696	3.646	3.642
PCA-t-stat	3.571	3.704	3.667	3.676
PCA-ΔR^2	3.568	3.704	3.671	3.676
PCA-lasso	3.566	3.710	3.666	3.668
PCA-ENet	3.567	3.707	3.666	3.676
Lasso	3.540	3.666	3.432	2.914
ENet	3.554	3.672	3.435	3.105
PCA-all	3.538	3.696	3.647	3.641
KS-t-stat	3.509	3.438	2.639	2.350
KS-ΔR^2	3.498	3.494	2.737	2.276
KS-lasso	3.477	3.583	3.344	−62.433
KS-ENet	3.515	3.596	3.322	−40.480

This table presents the portfolio performance, which is evaluated by the average realized utility (in percentage). h denotes the forecast horizon. In this portfolio exercise, we assume that a mean–variance investor will allocate her portfolio between WTI futures and risk-free bills by using different RV forecasts, which is based on a constant Sharpe ratio of 0.4 and a risk aversion coefficient of 2. The initial in-sample period is 1985:01-1997:12 and the out-of-sample evaluation period is 1998:01-2018:12.

values. This suggests that the investor is glad to pay more fees to use the PCA-VS models than the competing models. In summary, the PCA-VS models can help the investor realize the largest economic gains in a portfolio exercise.

6. Which predictors are selected?

The individual predictors are recursively selected by the supervised learning approaches of t-stat, ΔR^2, lasso, and ENet. It is pretty insightful to see how the pattern of VS evolves. To this end, we first depict the number of predictors recursively selected by t-stat, ΔR^2, lasso, and ENet each month throughout the out-of-sample period in Fig. 2. The average number of selected predictors for the 1-month horizon is much smaller than the ones for longer horizons. This echoes the findings for both in-sample and out-of-sample predictability. To be specific, a larger number of necessary predictors are selected for longer horizons, which can provide more useful and diverse forecasting information and thus contributes to stronger predictability over longer horizons. On average, the four VS methods select approximately 10 macroeconomic variables for the 1-month horizon. For the longer horizons (h = 3, 6, and 12), the t-stat and ΔR^2 method, on average, select 35 and 25 variables, respectively, and both the lasso and ENet select approximately 20 predictors. Overall, a relatively large number of predictors are selected. Another notable feature of Fig. 2 is that the number of selected predictors is relatively stable over time, typically in the fluctuation range of 25 predictors. Since the four learning approaches select variables based on their predictive power, Fig. 2 suggests that a sizable set of macro variables from the FRED-MD database is vital for forecasting oil futures market volatility.

Further, we explore the selection results for each predictor over the entire out-of-sample period. Fig. 3 shows the corresponding selection results. A selected (discarded) predictor is displayed as a black (white) bar each month. We can observe that the t-stat and ΔR^2 approaches are likely to select a considerable number of predictors from an identical category, showing a clustered selection pattern. For example, the two simple learning approaches simultaneously select more than 8 predictors from the category of interest and exchange rates (No. 81–102) for the 12-month horizon. Besides, for some cases, a relatively larger number of predictors from the categories of housing (No. 48–57) and consumption, orders, and inventories (No. 58–66) are also simultaneously selected. On the contrary, the lasso and ENet tend to select diverse predictors from different categories, showing a scattered selection pattern. This finding is in line with their statistical inference. Because of the L_1 parameter penalization, the lasso and ENet have a strong tendency to select one predictor from each category with highly correlated predictors. Looking at this figure carefully, we can detect that the selection pattern of the ENet is somewhat clustered relative to that of the lasso. This is because the ENet also introduces the L_2 penalty, resulting in a trade-off. It is noteworthy that the problem of selecting highly correlated predictors does not matter in this study because the PCA can alleviate the resulting issues of overfitting and multicollinearity.

Finally, to glean insight into how the supervised learning approaches select variable groups, Fig. 4 presents some heatmaps for the selection frequencies across eight variable groups. This figure first considers the entire out-of-sample period from 1998:01 to 2018:12 and then four sub-sample periods (each of the first three sub-samples covers 5 years, while the last includes 6 years). Consistent with the results in Fig. 2, a striking feature of Fig. 4 is that the selection frequencies for the t-stat and ΔR^2 methods are fairly higher than those of the lasso and ENet. There are two more interesting observations from the figure. First, the group of consumption, orders, and inventories show a high selection frequency for the short-term forecast horizon but a relatively low selection frequency for the long-term forecast horizon. This indicates that the variable group of consumption, orders, and inventories has a short-term impact on oil futures market volatility. Second, the selection frequency for the group of the stock market increases over time. This finding echoes the finding of Ma, Ji, et al. (2019) that a financial index, mainly consisting of stock market variables, shows the most powerful predictive ability for crude oil market volatility. Furthermore, Ma, Ji, et al. (2019) argue that oil financialization has been the key determinant for oil price behavior since the recent global financial crisis. Therefore, with the deepening of oil financialization, the variable group of the stock market shows an increasing role in forecasting oil market volatility. Similarly, a host of articles also find a close relationship between the oil and stock markets (see, e.g., Arouri et al., 2011, 2012; Chiang & Hughen, 2017; Creti et al., 2013; Degiannakis et al., 2018; Sim & Zhou, 2015; Wang et al., 2019).

Overall, the variable groups with relatively high selection frequency are stock market; housing; consumption, orders, and inventories; and interest and exchange rates. We also calculate the selection frequency for each predictor during the entire out-of-sample period. An overall selection frequency is further calculated by averaging individual selection frequencies in each panel (i.e., each

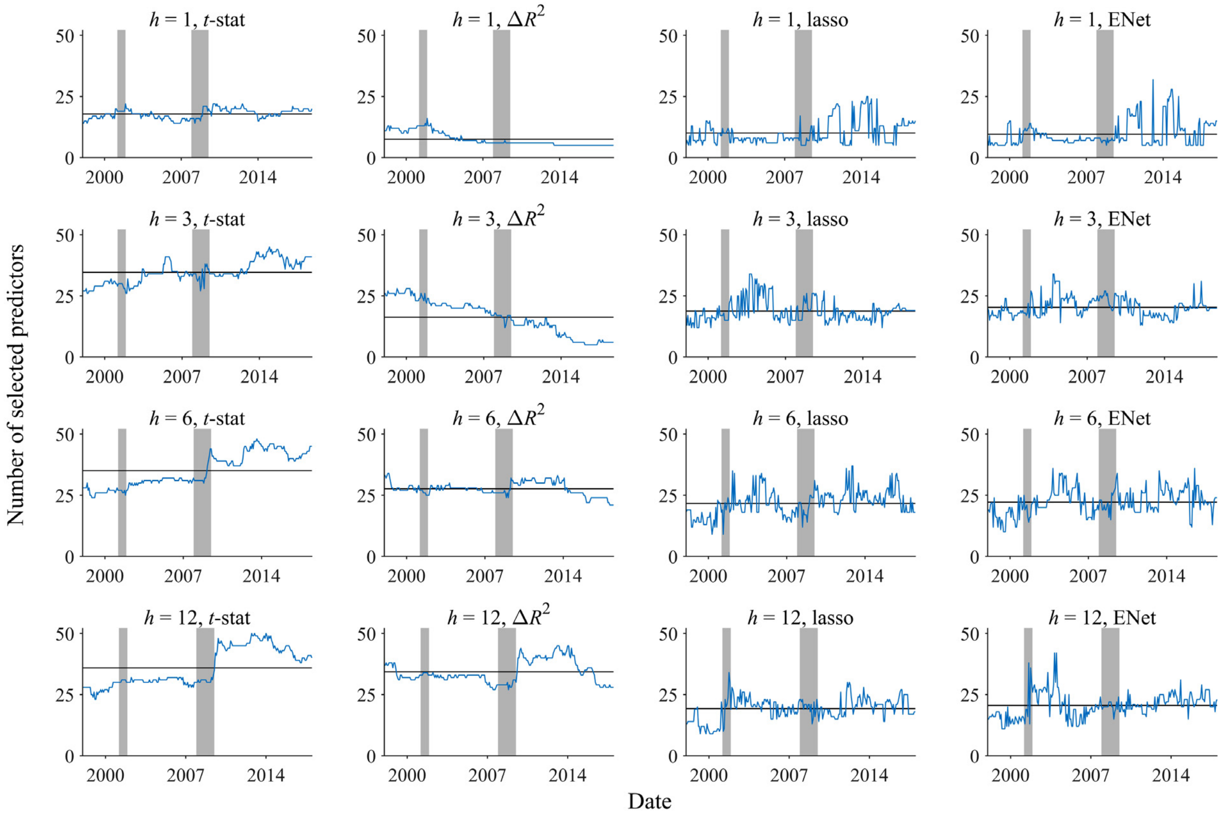

Fig. 2. The number of selected predictors.
This figure plots the number of predictors recursively selected by t-stat, ΔR^2, lasso, and ENet throughout the out-of-sample period from 1998:01 to 2018:12. h denotes the forecast horizon. The black solid horizontal line indicates the average number of selected predictors. Vertical bars delineate business-cycle recessions as dated by the NBER.

horizon and model) as shown in Fig. 3. The top five predictors with relatively high overall selection frequency are listed as follows:

1. VXOCLSx (CBOE S&P 100 Volatility Index): 98.4%.
2. NDMANEMP (All Employees: Nondurable goods): 74.2%.
3. S&P div yield (Dividend Yield for S&P's Composite Common Stock): 74.2%.
4. HOUSTS (Housing Starts, South): 68.9%.
5. T5YFFM (5-Year Treasury C Minus Effective Federal Funds Rate): 67.2%.

Notably, the VXO is the top selected predictor consistently for all four supervised learning approaches, resulting in a fairly high overall selection frequency of 98.4%. This figure is impressive as it is substantially larger than the second-highest selection frequency of 74.2%. This suggests that the option-implied volatility variable, VXO, is significantly more important than any other macroeconomic variable. This finding echoes a well-known empirical finding that there is a close link between the volatility behaviors of the oil and stock markets (see, e.g., Arouri et al., 2011, 2012; Creti et al., 2013; Degiannakis & Filis, 2017; Wang et al., 2018). In addition, this finding is also consistent with the well-recognized predictive ability of option-implied volatility. Many studies document that option-implied volatility is powerful for forecasting financial market volatility because option-implied volatility contains valuable information, such as investors' perception of the future asset return distribution, investor sentiment, investors' risk preference, and other information from informed investors (see, e.g., Busch et al., 2011; Jeon et al., 2020; Seo & Kim, 2015; Liang, Wei, & Zhang, 2020). Moreover, this finding also helps to understand why our recommended models show a stronger predictive power.

7. Conclusion

We attempt to improve the predictability of aggregate oil futures market volatility in a data-rich environment by employing machine learning approaches from the VS and common factor perspectives. More specifically, we use 127 macro variables from the FRED-MD database of McCracken and Ng (2016), facilitating "big-data" research in macroeconomic forecasting. We rely on four supervised learning approaches to select relatively useful predictors and thereby overcome the problem of overlapping information in such a large set of predictors, especially for the highly correlated ones in the same category. The first two traditional learning approaches are based on the significant regression slope with high-level t-statistic

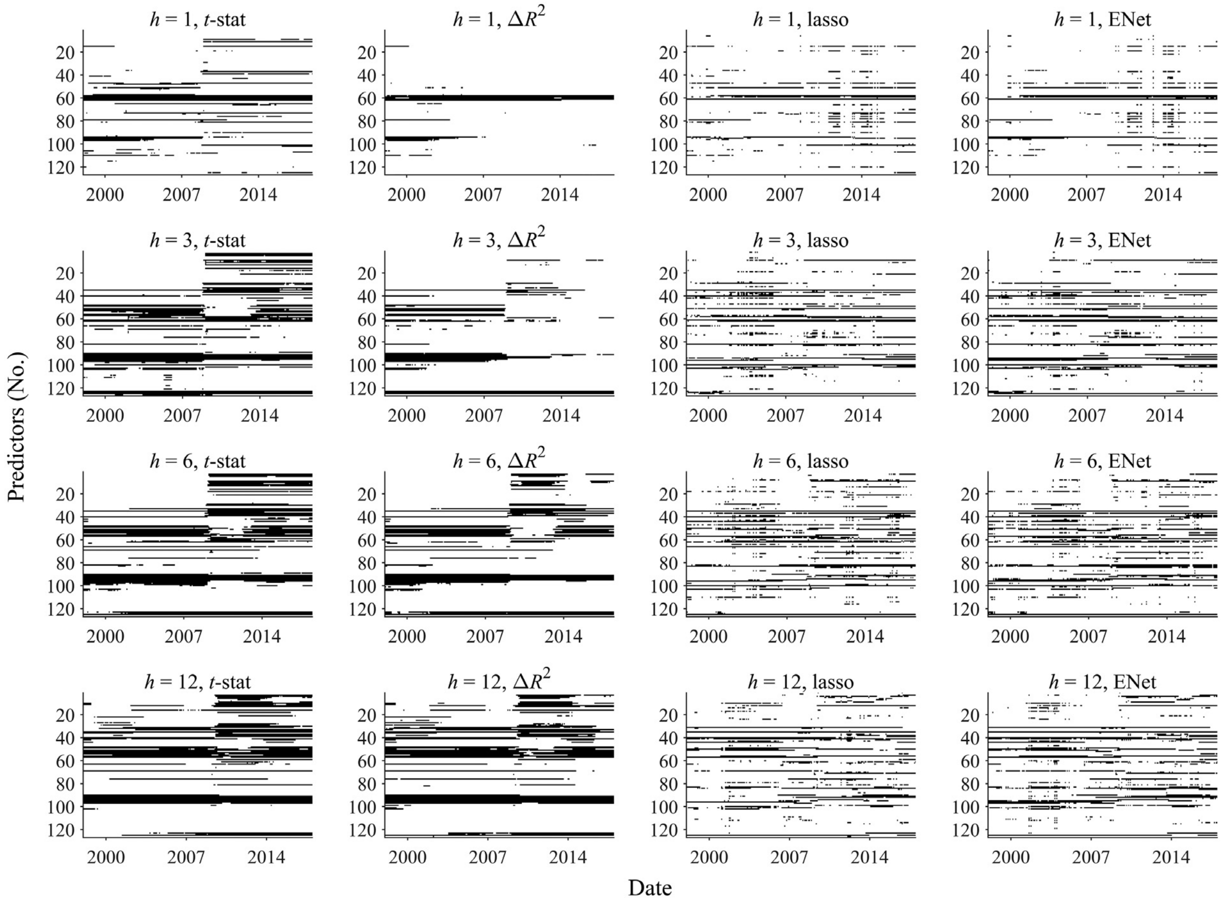

Fig. 3. Variable selection for each macro variable.
A selected (discarded) predictor is displayed as a black (white) bar each month. The predictors are recursively selected by t-stat, ΔR^2, lasso, and ENet throughout the out-of-sample period from 1998:01 to 2018:12. h denotes the forecast horizon.

(termed t-stat) and sizable incremental R^2 relative to the AR benchmark model (termed ΔR^2). The others are two popular shrinkage methods of the lasso and ENet. Further, we use the PCA technique to extract common factors based on the selected predictors, guarding against overfitting and multicollinearity.

We provide convincing evidence that the supervised PCA regression models (PCA-t-stat, PCA-ΔR^2, PCA-lasso, and PCA-ENet) outperform the prevailing AR benchmark and other related competing models both in-sample and out-of-sample. Our forecasting results are robust to alternative settings for VS, alternative AR model specifications, the choices of volatility target from the oil futures and spot markets, and many other situations. To allocate a portfolio between oil futures and risk-free bills, a mean–variance investor can realize utility gains using the supervised PCA regression models relative to the benchmark and competing models. The recommended models can also yield better hedging performance. The results document the economic value of our models, which is practically helpful for the participants in the crude oil market.

To shed light on which predictors matter for forecasting oil futures market volatility over time, we depict informative figures that provide a visual impression of the VS pattern. We observe that although VS shows a time-varying pattern, a relatively large number of predictors are recursively selected over the out-of-sample period. This indicates that many predictors are relevant to forecasting oil futures market volatility at different time points. Furthermore, all four supervised learning approaches are most likely to select the variables in the category of the stock market, especially the stock market volatility index (i.e., an option-implied volatility variable). This is in line with the well-documented empirical evidence regarding oil financialization and the closely associated volatility behaviors between the oil and stock markets. Given this, the participants in the crude oil market should pay attention to a large number of predictors, especially the stock market variables. Of course, valuable predictors would change over time, but option-implied volatility is always the most powerful.

Finally, we provide two interesting ideas for future research. First, since the conventional PCA technique is a simple dimension reduction method without any supervised learning, we could consider some modified dimension reduction methods with supervised learning, such as the partial least squares (PLS) approach (Kelly & Pruitt,

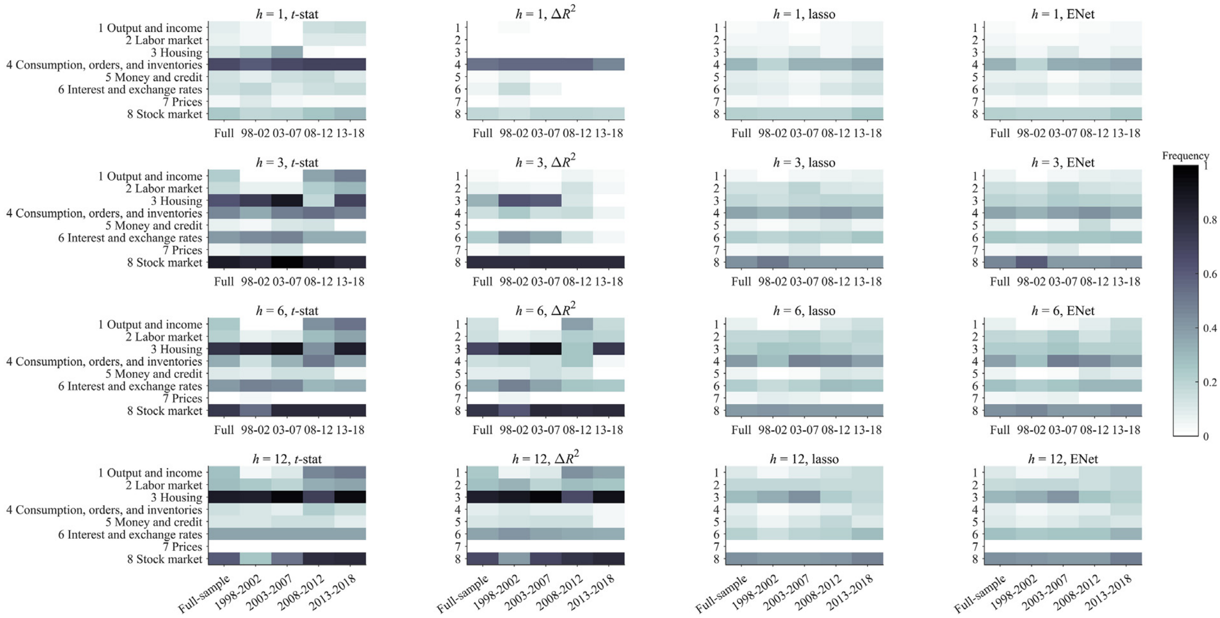

Fig. 4. Selection frequency for variable groups.
This figure depicts the heatmaps of the selection frequency for eight variable groups. The predictors are recursively selected by t-stat, ΔR^2, lasso, and ENet each month for various forecast horizons (h). The full-sample period here is the whole out-of-sample period from 1998:01 to 2018:12.

2015) and the scaled PCA (Huang et al., 2021). These new methods can probably improve the predictability of financial markets. Second, our machine learning approaches regarding VS could be used further to forecast combinations. For example, we can generate individual forecasts based on the selected predictors and average the forecasts. Alternatively, we can directly select accurate forecasts to combine by using our recommended machine learning approaches (see, e.g., Diebold & Shin, 2019).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful to Esther Ruiz (the editor), an anonymous associate editor, and two anonymous referees for insightful comments that significantly improved the paper. We also thank seminar participants at Changsha University of Science and Technology and Nanjing University of Science and Technology for helpful comments and suggestions. This work is supported by the National Natural Science Foundation of China [72001110, 71971191], the Fundamental Research Funds for the Central Universities [30919013232], and the Research Fund for Young Teachers of School of Economics and Management, NJUST [JGQN2009]. Yudong Wang acknowledges the financial support from the National Natural Science Foundation of China [71722015, 72071114].

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2021.12.013.

References

Arouri, M. E. H., Jouini, J., & Nguyen, D. K. (2011). Volatility spillovers between oil prices and stock sector returns: Implications for portfolio management. *Journal of International Money and Finance*, *30*, 1387–1405.

Arouri, M. E. H., Jouini, J., & Nguyen, D. K. (2012). On the impacts of oil price fluctuations on European equity markets: Volatility spillover and hedging effectiveness. *Energy Economics*, *34*, 611–617.

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

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

Busch, T., Christensen, B. J., & Nielsen, M. Ø. (2011). The role of implied volatility in forecasting future realized volatility and jumps in foreign exchange, stock, and bond markets. *Journal of Econometrics*, *160*, 48–57.

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

Chiang, I.-H. E., & Hughen, W. K. (2017). Do oil futures prices predict stock returns?. *Journal of Banking & Finance*, *79*, 129–141.

Christiansen, C., Schmeling, M., & Schrimpf, A. (2012). A comprehensive look at financial volatility prediction by economic variables. *Journal of Applied Econometrics*, *27*, 956–977.

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

Creti, A., Joëts, M., & Mignon, V. (2013). On the links between stock and commodity markets' volatility. *Energy Economics*, *37*, 16–28.

Degiannakis, S., & Filis, G. (2017). Forecasting oil price realized volatility using information channels from other asset classes. *Journal of International Money and Finance*, *76*, 28–49.

Degiannakis, S., Filis, G., & Arora, V. (2018). Oil prices and stock markets: a review of the theory and empirical evidence. *The Energy Journal*, *39*.

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

Diebold, F. X., & Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives. *International Journal of Forecasting*, *35*, 1679–1691.

Elder, J., & Serletis, A. (2010). Oil price uncertainty. *Journal of Money, Credit and Banking*, *42*, 1137–1159.

Ferderer, J. P. (1996). Oil price volatility and the macroeconomy. *Journal of Macroeconomics*, *18*, 1–26.

Fleming, J., Kirby, C., & Ostdiek, B. (2001). The economic value of volatility timing. *The Journal of Finance*, *56*, 329–352.

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

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

Goldstein, I., Spatt, C. S., & Ye, M. (2021). Big data in finance. *Review of Financial Studies*, *34*, 3213–3225.

Gong, X., & Lin, B. (2018). The incremental information content of investor fear gauge for volatility forecasting in the crude oil futures market. *Energy Economics*, *74*, 370–386.

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

Hamilton, J. D. (2009). Understanding crude oil prices. *The Energy Journal*, *30*.

Haugom, E., Langeland, H., Molnár, P., & Westgaard, S. (2014). Forecasting volatility of the US oil market. *Journal of Banking & Finance*, *47*, 1–14.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.

Hou, K., Mountain, D. C., & Wu, T. (2016). Oil price shocks and their transmission mechanism in an oil-exporting economy: A VAR analysis informed by a DSGE model. *Journal of International Money and Finance*, *68*, 21–49.

Huang, D., Jiang, F., Li, K., Tong, G., & Zhou, G. (2021). Scaled PCA: A new approach to dimension reduction. *Management Science*, http://dx.doi.org/10.1287/mnsc.2021.4020.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*, 297–307.

Jeon, B., Seo, S. W., & Kim, J. S. (2020). Uncertainty and the volatility forecasting power of option-implied volatility. *Journal of Futures Markets*, *40*, 1109–1126.

Jiang, F., Lee, J. A., Martin, X., & Zhou, G. (2019). Manager sentiment and stock returns. *Journal of Financial Economics*, *132*, 126–149.

Jo, S. (2014). The effects of oil price uncertainty on global real economic activity. *Journal of Money, Credit and Banking*, *46*, 1113–1135.

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

Kilian, L. (2009). Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *The American Economic Review*, *99*, 1053–1069.

Kilian, L., & Hicks, B. (2013). Did unexpectedly strong economic growth cause the oil price shock of 2003–2008? *Journal of Forecasting*, *32*, 385–394.

Kilian, L., & Park, C. (2009). The impact of oil price shocks on the US stock market. *International Economic Review*, *50*, 1267–1287.

Kilian, L., & Vigfusson, R. J. (2017). The role of oil price shocks in causing US recessions. *Journal of Money, Credit and Banking*, *49*, 1747–1776.

Koop, G., & Korobilis, D. (2011). UK macroeconomic forecasting with many predictors: Which models forecast best and when do they do so? *Economic Modelling*, *28*, 2307–2318.

Koop, G., & Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, *53*, 867–886.

Koop, G., & Korobilis, D. (2013). Large time-varying parameter VARs. *Journal of Econometrics*, *177*, 185–198.

Korobilis, D. (2013a). Bayeslan forecasting with highly correlated predictors. *Economics Letters*, *118*, 148–150.

Korobilis, D. (2013b). VAR forecasting using Bayesian variable selection. *Journal of Applied Econometrics*, *28*, 204–230.

Kristjanpoller, W., & Minutolo, M. C. (2016). Forecasting volatility of oil price using an artificial neural network-GARCH model. *Expert Systems with Applications*, *65*, 233–241.

Li, X., Shang, W., & Wang, S. (2019). Text-based crude oil price forecasting: A deep learning approach. *International Journal of Forecasting*, *35*, 1548–1560.

Liang, C., Wei, Y., Li, X., Zhang, X., & Zhang, Y. (2020). Uncertainty and crude oil market volatility: new evidence. *Applied Economics*, *52*, 2945–2959.

Liang, C., Wei, Y., & Zhang, Y. (2020). Is implied volatility more informative for forecasting realized volatility: An international perspective. *Journal of Forecasting*, *39*, 1253–1276.

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

Ludvigson, S. C., & Ng, S. (2009). Macro factors in bond risk premia. *Review of Financial Studies*, *22*, 5027–5067.

Ma, Y.-r., Ji, Q., & Pan, J. (2019). Oil financialization and volatility forecast: Evidence from multidimensional predictors. *Journal of Forecasting*, *38*, 564–581.

Ma, F., Liao, Y., Zhang, Y., & Cao, Y. (2019). Harnessing jump component for crude oil volatility forecasting in the presence of extreme shocks. *Journal of Empirical Finance*, *52*, 40–55.

Ma, F., Liu, J., Wahab, M. I. M., & Zhang, Y. (2018). Forecasting the aggregate oil price volatility in a data-rich environment. *Economic Modelling*, *72*, 320–332.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*, 802–808.

McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*, 574–589.

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.

Nonejad, N. (2017). Forecasting aggregate stock market volatility using financial and macroeconomic predictors: Which models forecast best, when and why? *Journal of Empirical Finance*, *42*, 131–154.

Pan, Z., Wang, Y., Wu, C., & Yin, L. (2017). Oil price volatility and macroeconomic fundamentals: A regime switching GARCH-MIDAS model. *Journal of Empirical Finance*, *43*, 130–142.

Paye, B. S. (2012). 'Déjà vol': PRedictive regressions for aggregate stock market volatility using macroeconomic variables. *Journal of Financial Economics*, *106*, 527–546.

Raftery, A. E., Kárný, M., & Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, *52*, 52–66.

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

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

Rapach, D., & Zhou, G. (2020). Time-series and cross-sectional stock return forecasting: New machine learning methods. In E. Jurczenko (Ed.), *Machine learning in asset management* (pp. 1–33).

Seo, S. W., & Kim, J. S. (2015). The information content of option-implied information for volatility forecasting with investor sentiment. *Journal of Banking & Finance*, *50*, 106–120.

Sévi, B. (2014). Forecasting the volatility of crude oil futures using intraday data. *European Journal of Operational Research*, *235*, 643–659.

Sim, N., & Zhou, H. (2015). Oil prices, US stock return, and the dependence between their quantiles. *Journal of Banking & Finance*, *55*, 1–8.

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

Wang, L., Ma, F., Liu, J., & Yang, L. (2020). Forecasting stock price volatility: New evidence from the GARCH-midas model. *International Journal of Forecasting*, *36*, 684–694.

Wang, Y., Pan, Z., Liu, L., & Wu, C. (2019). Oil price increases and the predictability of equity premium. *Journal of Banking & Finance*, *102*, 43–58.

Wang, Y., Wei, Y., Wu, C., & Yin, L. (2018). Oil and the short-term predictability of stock return volatility. *Journal of Empirical Finance*, *47*, 90–104.

Wang, Y., Wu, C., & Yang, L. (2013). Oil price shocks and stock market activities: Evidence from oil-importing and oil-exporting countries. *Journal of Comparative Economics*, *41*, 1220–1239.

Wei, Y., Liu, J., Lai, X., & Hu, Y. (2017). Which determinant is the most informative in forecasting crude oil market volatility: fundamental, speculation, or uncertainty?. *Energy Economics*, *68*, 141–150.

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

Zhang, Y., & Hamori, S. (2020). Forecasting crude oil market crashes using machine learning technologies. *Energies*, *13*(2440).

Zhang, Y., Ma, F., & Liao, Y. (2020). Forecasting global equity market volatilities. *International Journal of Forecasting*, *36*, 1454–1475.

Zhang, Y., Ma, F., & Wang, Y. (2019). Forecasting crude oil prices with a large set of predictors: Can LASSO select powerful predictors? *Journal of Empirical Finance*, *54*, 97–117.

Zhang, Y., Wei, Y., Zhang, Y., & Jin, D. (2019). Forecasting oil price volatility: Forecast combination versus shrinkage method. *Energy Economics*, *80*, 423–433.

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