Econometrics and Statistics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Econometrics and Statistics

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

Variable Selection in Macroeconomic Forecasting with Many Predictors

Zhenzhong Wang, Zhengyuan Zhu, Cindy Yu*

Department of Statistics, Iowa State University, 2438 Osborn Dr, Ames, 50011, IA, USA

ARTICLE INFO

Article history:
Received 5 October 2021
Revised 15 January 2023
Accepted 16 January 2023
Available online xxx

Keywords:
Best subset
Dimensional reduction
Factor augment model

ABSTRACT

In the data-rich environment, using many economic predictors to forecast a few key variables has become a new trend in econometrics. The commonly used approach is factor augment (FA) approach. This paper pursues another direction, variable selection (VS) approach, to handle high-dimensional predictors. VS is an active topic in statistics and computer science. However, it does not receive as much attention as FA in economics. This paper introduces several cutting-edge VS methods to economic forecasting, which includes: (1) classical greedy procedures; (2) l_1 regularization; (3) false-discovery-rate control methods, (4) gradient descent with sparsification and (5) meta-heuristic algorithms. Comprehensive simulation studies are conducted to compare their variable selection accuracy and prediction performance under different scenarios. Among the reviewed methods, a meta-heuristic algorithm called sequential Monte Carlo algorithm performs the best. Surprisingly the classical forward selection is comparable to it and better than other more sophisticated algorithms. In addition, these VS methods are applied on economic forecasting and compared with the popular FA approach. It turns out for employment rate and CPI inflation, some VS methods can achieve considerable improvement over FA, and the selected predictors can be well explained by economic theories.

1. Introduction

Recent development in information technology makes it possible to collect hundreds of economic variables in real time, with a reasonable cost. In such data-rich environment, using many economic predictors to forecast a few target variables has become a new trend in econometric research. In the last two decades, both theoretical and empirical works have been substantially built up on this direction, especially in the fields of macroeconomic forecasting [54,3] and real-time nowcasting [34,9]. This new trend has also made practical impact – economic forecasts using many predictors are currently being produced by fiscal and monetary authorities in both the U.S. [42,43] and Europe [26,33]. A key aspect of many-predictor forecasts is to impose suitable parsimonious structure on data so that the curse of dimensionality is circumvented and useful information can be extracted. There are two directions to accomplish this, which are based on two different assumptions about the economic data structure.

The first direction is factor augment (FA) approach. It has been found to produce superior forecasts over traditional methods such as AR and VAR, thus attained favor from both econometricians and practitioners. This approach assumes

* Corresponding author.
E-mail address: cindyyu@iastate.edu (C. Yu).

https://doi.org/10.1016/j.ecosta.2023.01.003

that many predictors are relevant to the target variable and they have a factor structure. Dynamic factor model is applied first to compress the information of predictors into a handful of estimated factors. Then, the factors are augmented to a linear forecasting equation for the target variable. The rationale behind this approach is that the common variation among many observed economic variables can be represented by a handful of unobserved factors, and disturbances to these factors correspond to the major aggregate shocks to the economy such as demand or supply shocks [56]. This idea has a long tradition in macroeconomics. One example is [54], which indicates the estimated factors can be interpreted as the diffusion indexes developed by NBER business cycle analysts to measure common movement in a set of macroeconomic variables.

Despite of the advantages of FA approach, there are a few drawbacks as well. First of all, it lacks explanation on the interrelationship among different economic variables, thus it cannot identify which predictors influence the target variable. Secondly, the estimated factors only capture the variation of major economic aggregates, but may lose information that is contained in a few predictors but beyond major economic aggregates. More importantly, the commonly used FA approach [54,55] does not take into account the target variable when estimating the factors, which means the factors used in the forecasting equation are the same no matter which target variable is being forecasted.

The second direction of many-predictor forecasts is to directly select the best predictors and their lagged values to carry out forecast, and we call it variable selection (VS) approach. This direction implies another rationale of the economic data – given the selected predictors in the forecasting equation, all others have insignificant prediction power to the target variable anymore. To be noticed, it does not mean the unpicked predictors are irrelevant or independent to the target variable. The forecasting equation derived from VS approach indicates that, conditional on the selected predictors, the remaining predictors have little prediction power on the target variable. VS is not a new topic, but it has not drawn as much attention as FA in economic forecasting. In contrast, VS has substantial development in other fields such as statistics, computer science, and bioinformatics, and impressive new methods and applications keep coming forward.

The first goal of this paper is to review several cutting-edge VS methods, and compare their performance with FA approach in the context of economic forecasting. One advantage of VS is its capability of interpreting the individual impact of each predictor on the target variable, including both direction and magnitude. This is helpful for understanding interrelationships among different economic variables. More importantly, VS can select predictors that may contain useful information beyond the aggregate economic activity explained by the factors in FA approach. Only including the important predictors will avoid overfitting issue, thus enhance prediction power. In our empirical studies in Section 5, we apply both FA and VS approaches to forecast three important macroeconomic variables – Employment (EMP), Industrial Production (IP) and Consumer Price Index (CPI), and find that some VS methods achieve considerable improvement over FA approach for EMP and CPI. Also the relationship between the target variable and the selected predictor can be well explained by economic theories.

The second goal of this paper is to evaluate several groups of VS methods, including both classical procedures and cutting-edge algorithms, in terms of both variable selection accuracy and out-of-sample forecasting. Due to the huge body of VS literature, it is impossible to do an exhaustive review for VS methodologies. For this paper, we only focus on the high dimensional regime (dimension of predictors is larger than the number of observations), which is the case of economic forecasting. We pick the following five groups of methods: (1) classical procedure (forward selection); (2) l_1 regularization (adaptive LASSO); (3) false-discovery-rate control methods; (4) gradient descent algorithms with sparsification (iterative hard thresholding and thresholding pursuit); and (5) a meta-heuristic algorithm called sequential Monte Carlo (SMC) proposed by [24]. All these VS methods are applied in the framework of linear regression. Their performance in both variable selection and out-of-sample forecasting are examined through several simulation studies. The results show that, SMC, the most time-consuming algorithm, works the best across all simulation settings. Surprisingly, the performance of the classical forward selection matches up to the SMC and better than other advanced modern methods (l_1 regularization and gradient descent algorithms with sparsification).

The remainder of the paper is structured as follows. Section 2 introduces the setting of economic forecasting with many predictors, and the implementation of FA and VS approaches. Section 3 briefly reviews the five groups of VS methods, including their methodologies, advantages and disadvantages. Several simulation studies are carried out in Section 4 to evaluate their performance. In Section 5 we apply these VS methods on economic forecasting using the FRED-MD database ([42]), and compares their forecasting performance with that of FA approach. Conclusions and discussions are presented in Section 6.

Notations Throughout this paper, bold letters denote vectors, unbold letters denote scalar quantities and calligraphy letters denote matrices. $\mathbf{0}$ and $\mathbf{1}$ stand for a vector of zeros and ones respectively. For a p-dimensional vector $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_p)'$, we use $\|\boldsymbol{\nu}\|_0 = \sum_{i=1}^p I(\nu_i \neq 0)$ with $I(.)$ being the indicator function, $\|\boldsymbol{\nu}\|_1 = \sum_{i=1}^p |\nu_i|$ and $\|\boldsymbol{\nu}\|_2 = \sqrt{\sum_{i=1}^p |\nu_i|^2}$ to denote the l_0 norm, l_1 norm and l_2 norm of $\boldsymbol{\nu}$ respectively. $\text{supp}(\boldsymbol{\nu})$ denotes the support of $\boldsymbol{\nu}$, i.e. $\text{supp}(\boldsymbol{\nu}) = \{\text{indices of nonzero elements in } \boldsymbol{\nu}\}$. For a set U, $|U|$ denotes its cardinality, i.e. the number of elements in U. $\boldsymbol{\nu}_U = [\nu_i]_{i\in U}$ stands for a sub-vector of $\boldsymbol{\nu}$ whose indices of elements belong to U. For every iterative algorithm, we use superscript (r) to stand for the r-th iteration.

2. Economic Forecasting with Many Predictors

In this section we first describe the setting of economic forecasting with many predictors, including notations and assumptions, then we outline FA and VS approaches under this framework.

2.1. Setting

We adopt the notations and assumptions per usual in economic forecasting literature [54,56,47]. Let y_{t+h}^h be the h-step ahead value of the variable to be forecasted. For example, in Section 5 we consider forecasts of 1, 3, 6 and 12-month growth of the Employment (EMP). Let EMP_t denote the value of EMP on month t. Then the h-month growth of EMP, at an annual rate, is

$$y_{t+h}^h = (1200/h)\log(EMP_{t+h}/EMP_t). \tag{1}$$

Let $\boldsymbol{Z}_t$ be the n-dimensional vector of predictor variables, which also includes the current value of the target variable y_t. In economic forecasting, both y_{t+h}^h and $\boldsymbol{Z}_t$ are required to be stationary. This is accomplished by suitable preliminary transformations which are determined by a combination of statistical tests and expert judgment. For instance, unit root tests indicate that the logarithm of industry production (IP) series (denoted as $\{IP_t\}$) has a unit root. Therefore, its appropriate transformation is taking the log first difference, i.e. the corresponding predictor variable is $z_t = \log(IP_t) - \log(IP_{t-1})$. After transformation, each predictor z_t is standardized to have mean zero and sample variance one. This standardization is required for FA approach and some VS methods.

2.1.1. FA Approach

FA approach assumes the predictor variables admit the following factor model representation:

$$\boldsymbol{Z}_t = \boldsymbol{\Lambda}\boldsymbol{F}_t + \boldsymbol{e}_t, \tag{2}$$

where $\boldsymbol{F}_t$ is the $s \times 1$ latent factors, $\boldsymbol{\Lambda}$ is the $n \times s$ matrix of factor loadings, and $\boldsymbol{e}_t$ is the vector of idiosyncratic components satisfying $E(\boldsymbol{e}_t|\boldsymbol{F}_t) = \boldsymbol{0}$ and finite second moments. Here the latent factors $\boldsymbol{F}_t$ are estimated by the principle component analysis. [54] has proved that the principal component estimator is point-wise (for any date t) consistent and has limiting mean squared error (MSE) over all t that converges to 0, under a suitable set of identifiability conditions. If some series contain missing values, the expectation-maximization (EM) algorithm described in [54] is utilized to estimate factors $\boldsymbol{F}_t$. After the factors have been estimated, the h-step ahead forecast is the linear projection of y_{t+h}^h onto the t-dated factors, y_t and their lagged values:

$$y_{t+h}^h = \alpha + \sum_{l=0}^{q-1}\alpha_l y_{t-l} + \sum_{l=0}^{m-1}\boldsymbol{\gamma}_l'\boldsymbol{f}_{t-l} + \varepsilon_{t+h}. \tag{3}$$

Here q is the auto-regressive order, m is the order of lagged factors and $\boldsymbol{f}_t$ is a vector of the first d factors in $\boldsymbol{F}_t$. In practice, q, m and d can be selected by some criteria, such as Schwarz's BIC [51] and forward cross-validation (FCV).

Note that factor model (2) only includes the current value of predictor variables ($\boldsymbol{Z}_t$) without considering their lagged values. The historical information of $\boldsymbol{Z}_t$ are incorporated to forecasting through the lagged value of factors ($\boldsymbol{f}_{t-l}$, $l = 1,\ldots,m-1$). In our empirical study, we have 128 monthly economic time series, thus the vector $\boldsymbol{Z}_t$ for FA approach is 128-dimensional. However, the predictors used in VS approach are $\boldsymbol{X}_t = \left(\boldsymbol{Z}_t', \boldsymbol{Z}_{t-1}', \ldots, \boldsymbol{Z}_{t-5}'\right)'$, including lagged values within half a year. Then the dimension of $\boldsymbol{X}_t$ is larger than the number of observations. In order to distinguish the predictors in these two different approaches, we use $\boldsymbol{Z}_t$ to denote the predictor variables in FA approach, and use $\boldsymbol{X}_t$ to denote the predictors in VS approach, respectively.

2.1.2. VS Approach

We apply various VS methods on economic forecasting through the following linear regression:

$$y_{t+h}^h = \beta_0 + \boldsymbol{X}_t'\boldsymbol{\beta} + \epsilon_{t+h}, \quad t = 1,\ldots,T, \tag{4}$$

where $\boldsymbol{X}_t = (X_{1t},\ldots,X_{pt})'$ is a p–dimensional vector of the predictors. In our real data application, $p = 6n = 768$. and $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)'$ is the vector of regression coefficients. The matrix form of (4) is as follows:

$$\boldsymbol{y} = \beta_0\boldsymbol{1} + \mathcal{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5}$$

$$\boldsymbol{y} = \left(y_{1+h}^h,\ldots,y_{T+h}^h\right)', \quad \mathcal{X} = [\boldsymbol{X}_1,\cdots,\boldsymbol{X}_T]' \quad \text{and} \quad \boldsymbol{\epsilon} = (\epsilon_{1+h},\ldots,\epsilon_{T+h})'. \tag{6}$$

Here we use $\boldsymbol{x}_i$ $(i = 1,\ldots,p)$ to denote each column of the model matrix $\mathcal{X}$, i.e. $\mathcal{X} = [\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p]$. As mentioned in Subsection 2.1, all $\boldsymbol{x}_i$'s are standardized to have mean zero and sample variance one.

The center part of the VS approach is the best subset problem with subset size k, which is given by the following optimization:

$$\min_{\boldsymbol{\beta}} \left\|\boldsymbol{y} - \beta_0\boldsymbol{1} - \mathcal{X}\boldsymbol{\beta}\right\|_2^2 \quad \text{subject to} \quad \left\|\boldsymbol{\beta}\right\|_0 \le k. \tag{7}$$

Here the l_0 norm of $\boldsymbol{\beta}$ (i.e. $\|\boldsymbol{\beta}\|_0$) counts the number of nonzeros in $\boldsymbol{\beta}$, which is bounded by k. Let $\hat{\boldsymbol{\beta}}_k$ be the optimal solution of (7), then the support of $\hat{\boldsymbol{\beta}}_k$, denoted as $\hat{U}_k := \{i : \hat{\beta}_i \neq 0\}$, is the best subset of predictors with size k. In practice, the subset size k can be determined by AIC [1], BIC, FCV or other criteria.

The discrete nature of cardinality constraint ($\|\boldsymbol{\beta}\|_0 \leq k$) poses a great difficulty in finding the global optimum. It requires comparison of all $\binom{p}{k}$ subsets of predictors, which is infeasible for large p. To the best of our knowledge, computing the optimal solution to problem (7) is in general deemed as intractable. However, the last few decades have seen a flurry of activity in developing algorithms trying to solve (7) at a reasonable time cost, with associated optimality under certain conditions. In Section 3, we will review five groups of VS methods that try to obtain the good sub-optimal solution more efficiently.

3. Overview of Variable Selection Methods

As there is a vast literature on this topic, we present a selective overview. We select the following five types of VS methods: (1) classical greedy procedures, (2) l_1 regularization methods, (3) false-discovery-rate control methods, (4) gradient descent algorithms with sparsification, and (5) meta-heuristic algorithms. The first three groups have already been investigated in many econometric literature, therefore their introduction are relatively concise. The last two groups are proposed in computer science and mathematical optimization but have not been widely adopted in economic forecasting. Thus these two groups will be introduced more elaborately. For each group, we mainly focus on the methods applied in our empirical study, presenting their ideas, advantages and disadvantages. The algorithm details can be found in Appendix B.

3.1. Classical Greedy Procedures

Classical VS procedures such as forward selection (FS), backward elimination (BE), and stepwise regression (SR) are available in many statistical software packages. These algorithms are greedy algorithms, which follow the heuristic of making the locally optimal choice at each iteration with the intent of finding a global optimum. For example, when adding a new predictor to the model, FS selects the one which maximize the decrement of sum of square errors (SSE) given the predictors already included in the model, until the model has k predictors in total. As a result, such greedy search only examines a small portion of possible subsets of predictors and may be trapped in a local solution thus cannot guarantee to obtain the global optimum.

Due to the nature of greedy algorithms, the classical procedures seem to be inferior to modern VS algorithms such as LASSO. However, the latter also rely on certain assumptions to achieve global optimality, and these assumptions are hard to be verified in practice. In addition, FS and BE do not involve any tuning parameters, while more advanced algorithms are usually sensitive to the choice of tuning parameters and initial values. Thus, it is meaningful to compare classical greedy procedures with the modern algorithms in terms of their empirical performance. Moreover, another advantage of FS and BS is their fast update of model estimation when adding or deleting a predictor, which can avoid calculating the inverse of $\mathcal{X}'\mathcal{X}$. This fast updating algorithm and the detailed procedures of FS and BE are presented in Appendix B.2.

3.2. l_1 Regularization Methods

Since it is hard to directly solve the l_0 constraint optimization (7), a group of methods pursue another direction of VS – l_1 regularization, which is a convex relaxation of the l_0 constraint. One representative is LASSO [57]:

$$\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \beta_0 \mathbf{1} - \mathcal{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}. \tag{8}$$

As a surrogate for problem (7), l_1 regularized methods produce a sparse estimation by shrinking many coefficients toward zero. There has been a large amount of work on this topic in terms of algorithms, theoretical properties and real world applications. Readers can refer to the books [17], [36], [60] and references therein.

l_1 regularization methods enjoy several attractive properties. The first advantage of l_1 regularization, which is also an important reason to its popularity, is great computational efficiency. The problem (8) is a convex quadratic optimization and there are several efficient algorithms to solve it. For example, pathwise coordinate optimization [29] can compute the solution path at the same cost as a least squares calculation. Second, under some conditions it can be shown that LASSO can recover the true sparseness of $\boldsymbol{\beta}$ (i.e. consistent variable selection). However, one main and restrictive condition to achieve this, the so-called neighborhood stability condition, is related to the model matrix $\mathcal{X}$ and difficult to be verified in practice. Additionally, the magnitudes of non-zero coefficients cannot be too small. To achieve consistent variable selection, the smallest non-zero coefficient must satisfy the following beta-min condition:

$$\inf_{\beta_j \neq 0} |\beta_j| \gg \sqrt{s_0 log(p)/T}, \tag{9}$$

where s_0 is the number of non-zero coefficients in the true model.

As argued in [27], the original LASSO (8) leads to biased coefficient estimates. To address this shortcoming, several approaches such as SCAD [27], adaptive LASSO [67] and minimax convex penalty [65] are proposed. It has been shown that their estimates possess the so-called oracle properties: (1) identifies the true set of predictors asymptotically; (2) has the

optimal convergence rate. The adaptive LASSO has the following objective function:

$$\min_{\boldsymbol{\beta}} \left\{ \left\| \boldsymbol{y} - \beta_0 \mathbf{1} - \mathcal{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \sum_{i=1}^{p} w_i |\beta_i| \right\}. \tag{10}$$

Here, w_i $(i = 1, \ldots, p)$ is the penalty weight of β_i that can be derived from $\sqrt{T}$-consistent estimator such as LASSO estimator $(w_i = 1/|\hat{\beta}_i^{LASSO}|)$ or ridge estimator $(w_i = 1/|\hat{\beta}_i^{ridge}|)$.

Besides the above methods, other LASSO-based approaches have been investigated in recent decades. [68] point out that LASSO tends to select only one predictor within a group of highly correlated predictors and performs worse than ridge regression when the number of predictors p is larger than number of observations T. To solve these problems, they proposed elastic net by adding an additional l_2 penalty to optimization problem (8). Their approach allows to select strongly correlated predictors together and improves forecasting performance when $p \gg T$. [58] propose fused LASSO which penalizes the l_1-norm of both the coefficients and their successive differences. Thus it encourages both sparsity of the coefficients and forces the coefficients to vary smoothly. [64] propose group LASSO which guarantees pre-defined groups of covariates to be in or out of the model together. This approach is suitable for scenarios in which the predictors belong to pre-defined groups, for example, a collection of dummy variables representing the levels of a categorical predictor. Further extensions of group LASSO methods includes sparse group LASSO [49] which can select individual covariates within a group, and overlap group LASSO [37,66] which allows covariates to be shared across groups.

LASSO-based approaches also have been investigated in macroeconomic forecasting by many authors. [23] applies a Bayesian framework of LASSO on forecasting of Industrial Production and Consumer Price Index. Targeted FA approach proposed by [3] uses least angle regression to select a subset of predictors, from which the dynamic factors are constructed. [41] compare LASSO, elastic net and group LASSO with FA approach [54] on forecasting of 20 macroeconomic variables and show that LASSO-type approaches perform better in out-of-sample forecast. In addition, there are quite a few papers investigating LASSO-type penalty in autoregressive-moving-average model and vector autoregressive model, for example [61] [7] and [48], which are out of the scope of this paper.

In spite of the good properties and wide usage of LASSO methods mentioned above, l_1 regularization is a relaxation of VS problem, thus do not provide provably optimal solution to (7). [12] indicates that the upper bound of l_2 estimator error from l_0 regularization (7) will be much smaller than that from the LASSO estimator, if the pairwise correlations between the predictors are quite high. In our simulation studies, l_1 regularization tends to introduce a lot of spurious variables. As a consequence, it may be inferior to other types of methods in terms of identifying the important predictors.

3.3. False Discovery Rate Control

To solve the problem of over fitting, quite a few approaches have been proposed to control the false discovery rate (FDR). This idea comes from the breakthrough work of [10] in the framework of large scale hypothesis testing. Nowadays FDR control becomes an important criterion of statistical inference in large scale hypothesis testing and various modifications of [10] procedure have been proposed. Another group of FDR-control methods is knockoff filter [5] and its extensions such as [18] and [6]. In this paper, we choose knockoff filter as a representative for the FDR-control methods.

Consider the high-dimensional linear regression problem (5). Knockoff filter constructs a matrix $\tilde{\mathcal{X}}$ which mimics $\mathcal{X}$ without looking at $\boldsymbol{y}$. It satisfies a subtle pairwise exchangeable condition: for any j, the joint distribution of the augmented random matrix $[\mathcal{X}, \tilde{\mathcal{X}}]$ does not change if its jth and $(j+p)$th columns are swapped. The matrix $\tilde{\mathcal{X}}$ is called knockoff matrix of $\mathcal{X}$. By its construction, $\tilde{\mathcal{X}}$ have the same distribution as $\mathcal{X}$ but it is known to be a null set, that is, $\tilde{\mathcal{X}}$ is not important for $\boldsymbol{y}$. So $\tilde{\mathcal{X}}$ serves as a "negative control" for $\mathcal{X}$: if a predictor $\boldsymbol{x}_j$ belongs to the true model, it should has larger prediction power than its knockoff copy $\tilde{\boldsymbol{x}}_j$. VS approach is applied on the augmented set $Z := [\mathcal{X}, \tilde{\mathcal{X}}]$. Then, the importance of predictor $\boldsymbol{X}_j$ can be detected by comparing its predictive power to that of its knockoff copy $\tilde{\boldsymbol{X}}_j$ $(j = 1, 2, \ldots, p)$. The knockoff procedure picks only variables which are better than their knockoff copies. Take LASSO for example, which has the following optimization:

$$\min_{\boldsymbol{\beta}, \boldsymbol{\gamma}} \left\{ \left\| \boldsymbol{y} - \beta_0 \mathbf{1} - \mathcal{X}\boldsymbol{\beta} - \tilde{\mathcal{X}}\boldsymbol{\gamma} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right\}. \tag{11}$$

For any given penalty parameter λ, the prediction power of $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$ can be represented by their coefficient estimator, $\hat{\beta}_j(\lambda)$ and $\hat{\gamma}_j(\lambda)$, respectively. Then a statistics is constructed to compare $\hat{\beta}_j(\lambda)$ and $\hat{\gamma}_j(\lambda)$:

$$W_j = |\hat{\beta}_j(\lambda)| - |\hat{\gamma}_j(\lambda)|, \quad j = 1, 2, \ldots, p. \tag{12}$$

Clearly, $W_j > 0$ means X_j is more important than its knockoff copy $\tilde{\boldsymbol{X}}_j$. At last, the predictors with W_j's greater than a threshold M are selected. The threshold M is defined as:

$$M = \min \left\{ m : \frac{1 + \#\{W_j \le -m\}}{\#\{W_j > m\}} \le \alpha \right\}, \tag{13}$$

where α is the target FDR.

The most significant challenge of implementing knockoffs is how to constructing $\tilde{\mathcal{X}}$ since it requires nontrivial knowledge on the high-dimensional distribution of $\mathcal{X}$. Several algorithms have been proposed, for example [18], [8] and [50]. In this paper, we use the R package `knockoff` to implement the the knockoff procedure.

3.4. Gradient Decent Algorithms with Sparsification

In the fields of computer science and signal processing, a group of algorithms, referred to gradient decent with sparsification (GDS) algorithms, are developed to directly provide a good solution to the l_0-constraint optimization (7). As the name suggests, these methods are extensions of gradient descent algorithms which impose sparsity on the coefficient estimate. These algorithms include but not limited to iterative hard thresholding [14], compressive sampling matching pursuit [46], subspace pursuit [22], hard thresholding pursuit [28] and orthogonal matching pursuit with replacement [38]. The analysis of their theoretical properties in high dimensional regression setting can be found in [39], [13] and references therein. In general, this group of methods aim to minimize a loss function $f(\boldsymbol{\beta})$ subject to the l_0 constraint:

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_0 \leq k. \tag{14}$$

In our case, $f(\boldsymbol{\beta})$ equals to $\|\boldsymbol{y} - \beta_0\mathbf{1} - \mathcal{X}\boldsymbol{\beta}\|_2^2$, the SSE of linear regression. Without considering the l_0 constraint, gradient descend algorithm minimizes $f(\boldsymbol{\beta})$ iteratively by updating $\boldsymbol{\beta}$ as:

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} - \eta \nabla f(\boldsymbol{\beta}^{(r)}), \tag{15}$$

where superscript (r) stands for the rth iteration, and $\nabla f(\boldsymbol{\beta}^{(r)}) = 2\mathcal{X}'\mathcal{X}\boldsymbol{\beta}^{(r)} - 2\mathcal{X}'\boldsymbol{y}$ is the gradient of $f(\boldsymbol{\beta})$ at $\boldsymbol{\beta}^{(r)}$, and η is the step size. Since a k-sparse vector is desired, GDS-type algorithms modify the updating equation (15) in some ways so that $\boldsymbol{\beta}^{(r+1)}$ becomes a k-sparse vector. In the following, we will introduce four GDS-type algorithms.

Iterative hard thresholding (IHT), also known as projected gradient descent, modifies the updating equation by adding a hard thresholding operator $H_k(.)$ to it:

$$\text{IHT:} \quad \boldsymbol{\beta}^{(r+1)} = H_k\left(\boldsymbol{\beta}^{(r)} - \eta \nabla f\left(\boldsymbol{\beta}^{(r)}\right)\right). \tag{16}$$

For any vector $\boldsymbol{v}$, the hard thresholding operator $H_k(\boldsymbol{v})$ keeps the k largest (in magnitude) elements of $\boldsymbol{v}$ and set the rests to zero. This operator is the simplest way to obtain a $k-$sparse vector. It has been also proven that for any arbitrary vector $\boldsymbol{v}$, $H_k(\boldsymbol{v})$ is the closest k-sparse vector to it in l_2 distance.

IHT directly converts $\boldsymbol{\beta}^{(r)} - \eta \nabla f(\boldsymbol{\beta}^{(r)})$ into a sparse vector, while some other methods, such as compressive sampling pursuit (CoSaMP) and subspace pursuit (SP), chase a good support of $\boldsymbol{\beta}$ based on the gradient descent and then find the best fit within this support. Specifically, the CoSaMP algorithm iterates the following three-step scheme:

$$\text{CoSaMP:} \quad U^{(r+1)} = \text{supp}\left(\boldsymbol{\beta}^{(r)}\right) \cup \left\{\text{indices of } 2k \text{ largest elements of } \nabla f\left(\boldsymbol{\beta}^{(r)}\right)\right\}, \tag{17}$$

$$\tilde{\boldsymbol{\beta}}^{(r+1)} = argmin_{\text{supp}(\boldsymbol{\beta}) \subseteq U^{(r+1)}} \|\boldsymbol{y} - \beta_0\mathbf{1} - \mathcal{X}\boldsymbol{\beta}\|_2^2 \tag{18}$$

and

$$\boldsymbol{\beta}^{(r+1)} = H_k\left(\tilde{\boldsymbol{\beta}}^{(r+1)}\right). \tag{19}$$

SP algorithm is similar to CoSaMP, except that it replaces $2k$ with k in the first step of CoSaMP and modifies the third step:

$$\text{SP:} \quad U^{(r+1)} = \text{supp}\left(\boldsymbol{\beta}^{(r)}\right) \cup \left\{\text{indices of } k \text{ largest elements of } \nabla f(\boldsymbol{\beta}^{(r)})\right\}, \tag{20}$$

$$\tilde{\boldsymbol{\beta}}^{(r+1)} = argmin_{\text{supp}(\boldsymbol{\beta}) \subseteq U^{(r+1)}} \|\boldsymbol{y} - \beta_0\mathbf{1} - \mathcal{X}\boldsymbol{\beta}\|_2^2, \tag{21}$$

$$V^{(r+1)} = \left\{k \text{ largest elements of } \tilde{\boldsymbol{\beta}}^{(r+1)}\right\}$$

and

$$\boldsymbol{\beta}^{(r+1)} = argmin_{\text{supp}(\boldsymbol{\beta}) \subseteq V^{(r+1)}} \|\boldsymbol{y} - \beta_0\mathbf{1} - \mathcal{X}\boldsymbol{\beta}\|_2^2. \tag{22, 23}$$

Unlike IHT, CoSaMP and SP have one and two OLS calculations respectively. Since OLS is the most time consuming part in each iteration, these two methods have more computational cost than IHT within one iteration. However, on the other hand, OLS offers the best fit within the proposed support, which may make these two methods converge with less iterations.

[28] combines the hard thresholding operator in IHT and the idea of pursuing a good support of $\boldsymbol{\beta}$ (CoSaMP and SP) into the following hard thresholding pursuit (HTP) algorithm:

$$\text{HTP:} \quad \tilde{\boldsymbol{\beta}}^{(r+1)} = H_k\left(\boldsymbol{\beta}^{(r)} - \eta \nabla f(\boldsymbol{\beta}^{(r)})\right) \tag{24}$$

and

$$\boldsymbol{\beta}^{(r+1)} = argmin_{\text{supp}(\boldsymbol{\beta}) \subseteq \text{supp}\left(\tilde{\boldsymbol{\beta}}^{(r+1)}\right)} \left\|\boldsymbol{y} - \beta_0 \mathbf{1} - \mathcal{X}\boldsymbol{\beta}\right\|_2^2. \tag{25}$$

The difference between HTP and CoSaMP/SP lies on their ways of proposing the support of $\boldsymbol{\beta}$. HTP uses the support of IHT result while CoSaMP and SP derives the support by the result of previous iteration $\boldsymbol{\beta}^{(r)}$ and the corresponding gradient $f(\boldsymbol{\beta}^{(r)})$.

In general, GDS-style algorithms are not limited to the aforementioned four algorithms, and there is no guarantee that one algorithm outperforms the others. In this paper, we select IHT and HTP as representatives of this group of methods. The initial input $\boldsymbol{\beta}^{(0)}$ should be a k-sparse vector, typically $\boldsymbol{\beta}^{(0)} = \mathbf{0}$. The iteration is stopped when the difference between $\boldsymbol{\beta}^{(r+1)}$ and $\boldsymbol{\beta}^{(r)}$ is small enough or the maximum number of iterations is reached.

3.5. Sequential Monte Carlo

Optimization (7) is a combinatory optimization problem whose search space is discrete. Some meta-heuristic algorithms, such as simulated annealing [40,19] and genetic algorithms [35], are dedicated to solve such combinatory optimization, and become appealing to variable selection [20,16,15]. Unlike greedy algorithms that always reject worse solutions, meta-heuristic algorithms accept worse solution with a probability to yield a more extensive research. In this subsection, we introduce one meta-heuristic algorithm proposed by [24], called sequential Monte Carlo (SMC). This algorithm incorporates the idea of simulated annealing and Monte Carlo methods to considerably extend the searching space. Even the key idea involves Monte Carlo, SMC algorithm is different from the Bayesian variable selection methods that assume hierarchical distribution for the data and variables are selected by imposing some spike-and-slab priors on the model coefficients. Indeed, SMC approach is not a Bayesian method. It does not require any distributional assumption and solves the l_0 constraint optimization (7) in its original form.

For any nonempty set of indices U, let $\mathcal{X}_U = [\boldsymbol{x}_i]_{i\in U}$ be the sub-matrix of $\mathcal{X}$ whose columns belong to set U, and let $\boldsymbol{\beta}_U = [\beta_i]_{i\in U}$ be the corresponding sub-vector of coefficients. Then the l_0 constraint optimization (7) is rewritten as a minimization problem:

$$\min_{|U|=k} \left\|\boldsymbol{y} - \hat{\beta}_0 \mathbf{1} - \mathcal{X}_U \hat{\boldsymbol{\beta}}_U\right\|_2^2, \tag{26}$$

where $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}_U$ are the OLS estimates corresponding to the intercept and $\mathcal{X}_U$. Let $\mathbb{U}_k$ be the collection of all the permutations of k indices, i.e. $\mathbb{U}_k = \{U \subseteq \{1, \cdots, p\} : |U| = k\}$. [24] assigned a discrete distribution function $f(U)$ on $\mathbb{U}_k$:

$$f(U) \propto \exp\left\{-\left\|\boldsymbol{y} - \hat{\beta}_0 \mathbf{1} - \mathcal{X}_U \hat{\boldsymbol{\beta}}_U\right\|_2^2\right\}. \tag{27}$$

Then finding the global optimum of (26) is done through generating a representative sample from this distribution. Obviously, among all permutations in $\mathbb{U}_k$, the optimal one corresponds to the peak of this distribution thus is more likely to be generated. Distribution f is defined over permutations instead of combinations because permutations are easier to sample. Since this distribution has no tractable analytical solution and is multimodal, it is hard to directly construct a suitable proposal distribution for it. SMC applies the idea of distribution tempering that starts from an easy-to-sample distribution f_0, moves "smoothly" to the complex target distribution f by composing a sequence of artificial intermediate distributions:

$$f_0, f_1, \ldots, f_J, \quad f_j(U) \propto f^{\gamma_j}(U) f_0^{1-\gamma_j}(U). \tag{28}$$

The distribution sequence $\{f_j\}_{j=1}^J$ is called "distribution-tempering bridge" and sequence $\{\gamma_j\}$ satisfies $0 = \gamma_0 < \gamma_1 < \cdots < \gamma_J = 1$. Clearly, $\gamma_0 = 0$ corresponds to the initial distribution f_0, and $\gamma_J = 1$ corresponds to target distribution f. For each round j, the choice of γ_j is self adapted by the algorithm, which makes the difference between f_{j-1} and f_j small enough so that the former distribution, f_{j-1}, can be a good proposal for the latter distribution, f_j.

The initial distribution, f_0, takes into consideration the prediction power of each individual predictor. Let R_i^2 be the R^2 of a single linear regression with the i-th predictor ($i = 1, \ldots, p$). The initial distribution, f_0, is the random sampling without replacement based on inclusion probability $q_i = R_i^2 / \left(\sum_{i=1}^p R_i^2\right)$:

$$f_0(\{i_1, i_2, \cdots, i_k\}) = q_{i_1} \frac{q_{i_2}}{1 - q_{i_1}} \frac{q_{i_3}}{1 - q_{i_1} - q_{i_2}} \cdots \frac{q_{i_k}}{1 - q_{i_1} - \cdots - q_{i_{k-1}}}, \tag{29}$$

for any $\{i_1, i_2, \cdots, i_k\} \in \mathbb{U}_k$. After obtaining the initial sample, denoted by $\Omega_0 = \{U_{0,m}\}_{m=1}^M$ with M being the sample size, we can generate a representative sample for the next distribution f_1 by following the three-step scheme:

- **Reweighting:** First, an importance weight is assigned to each $U_{0,m}$, which is the ratio between the probability of $U_{0,m}$ under distribution f_1 and that under f_0, i.e. $w_{1,m} \propto f_1(U_{0,m})/f_0(U_{0,m})$. The parameter γ_1 in f_1 is determined to guarantee the effective sample size implied by these weights not smaller than a threshold, say $M/2$.
- **Resampling:** Use the importance weights to resample the $U_{0,m}$'s in Ω_0 which will result a new sample satisfying the distribution f_1.
- **Support boosting:** After resampling, some $U_{0,m}$'s are duplicated to reflect their relatively high importance weights while some are excluded due to their low weights. Thus the empirical support (distinct U's in the sample) is shrunk. Boosting the empirical support is accomplished by several Metropolis Hastings moves which reduce duplicates, enlarge the number of distinct members, and retain distribution f_1 for the sample at the meantime. After this step, we will get a representative sample for f_1, denoted as $\Omega_1 = \{U_{1,m}\}_{m=1}^{M}$.

By repeating this three-step scheme until γ reaches 1, we finally arrive at a representative sample for the target distribution f. At last, a k-fold duplication technique [25] is carried out to enlarge the sample size k times, and the best subset of predictors is the U with the smallest SSE in the final sample. The detailed algorithm can be found in Appendix B.3, and readers can refer to [24] for its theoretical properties.

4. Simulation Studies

We present a variety of computational experiments to: (1) evaluate different types of VS methods in terms of variable selection accuracy and out-of-sample prediction; (2) compare FA approach to VS approach under the framework of time series forecasting. These two goals are addressed in Section 4.1 and 4.2 respectively.

4.1. Evaluation of Different Types of VS Methods

We conduct three simulation studies to compare among different VS methods in terms of variable selection accuracy and out-of-sample prediction. For each group of VS methods, we select one or two representatives: (1) FS; (2) adaptive LASSO (adaLASSO); (3) knockoff; (4) IHT and HTP algorithms; and (5) SMC algorithm. In our simulation studies, we investigate the $p > T$ case, thus BE is not applicable and excluded. In addition, variable screening may be also helpful for variable selection. We choose generalized least squares screening (GLSS) with adaLASSO (denoted as GLSS+adaLASSO) proposed by [63] in both simulation studies and real data application. It contains two stages: First GLSS is used to screen out less important predictors, then adaLASSO is applied on the screened predictors for the final variable selection. GLSS in the first stage takes temporal dependence into considerations thus may be suitable for time series forecasting. In order to examine their performance, we consider five criteria:

- precision = True Positive/(True Positive + False Positive);
- recall = True Positive/(True Positive + False Negative);
- dice coefficient (DC) = 2 True Positive/(2 True Positive + False Positive + False Negative);
- mean squared prediction error (MSPE) on test set (sample size=100);
- time cost in minutes for one simulation.

Here, "True Positive" stands for the number of predictors that are correctly identified, "False Positive" is the number of predictors which are wrongly selected, and "False Negative" counts the number of predictors belonging to the true model but missed by the VS method. In this way, recall is the proportion of predictors in the true model that correctly identified, while precision is the proportion of the selected predictors that are truly significant. These two criteria quantify different aspects of the variable selection accuracy, and higher value indicates higher accuracy. However, there is a trade-off between precision and recall: as more predictors are selected, recall tends to increase while precision would decrease. The criteria, DC, can balance this trade-off and serve as an overall measurement of variable selection accuracy. Since this paper aims to apply VS methods on economic forecasting, we also include the MSPE as the criterion to evaluate out-of-sample prediction.

4.1.1. Simulation Settings

Simulation is carried out under the linear regression framework. The true model is:

$$y_t = \sum_{j=1}^{K} \beta_j x_{jt} + \epsilon_t, \quad \epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2), \quad t = 1, \ldots, T. \tag{30}$$

The regressors are selected from p potential predictors with $p \gg K$. All the potential predictors are generated from normal distribution with mean zero and variance one. Note that in this subsection we do not consider temporal dependence thus the observations are independent across time. In total, we construct three settings to evaluate VS methods under different scenarios. The first one adopts the simulation setting of [24], which presents moderate high dimensional case ($p = 900$ and $T = 200$). The other two settings are cases of ultra-high dimension ($p = 2000$, $T = 50,\ 100,\ 200$), with setting 2 imposing independent structure and setting 3 imposing correlated structure among predictors.

Table 1
Five quantiles of the number of selected predictors by each VS method and the mean of in-sample R^2 among 100 repetitions for setting 1.

	FS	SMC	adaLASSO	IHT	HTP	knockoff	GLSS+adaLASSO
Theoretical $R^2 = 0.8$							
Min.	5	5	16	6	5	0	3
1st Qu.	7	7	32	9	8	4	9
Median	8	7	38	10	8	8.5	11
3rd Qu.	9	8	46	11	10	18	15
Max.	13	12	63	17	14	62	30
R^2	0.80	0.80	0.85	0.76	0.78	0.64	0.44
Theoretical $R^2 = 0.5$							
Min.	5	5	13	5	5	0	3
1st Qu.	7	7	26	7	7	2	9
Median	7	7	34	7	7	8	11
3rd Qu.	7	8	42	8	7	16	15
Max.	11	12	83	13	11	67	30
R^2	0.55	0.57	0.64	0.51	0.52	0.41	0.44

- **Setting 1:** 900 potential predictors are equally divided into three groups with 300 in each. The correlation within each group are 0.1, 0.4 and 0.8 respectively, and predictors across different groups are independent. Within each group, four predictors are included in the true model with coefficients 0.1, 0.4, 0.7 and 1. In total there are $K = 12$ relevant predictors. To ascertain the impact of signal strength in variable selection, two levels of theoretical R^2 (0.8 and 0.5) are considered, which determine the values of σ^2.
- **Setting 2:** 2000 potential predictors are generated independently from $N(0, 1)$. The true model only includes five of them ($K = 5$) and all of their coefficients are one. The theoretical R^2 is set to be 0.8. Three different sample sizes are considered (T =50, 100 and 200) to assess its influence on model performance.
- **Setting 3:** 2000 potential predictors are divided into four groups with 500 in each. The correlation within each group are 0.1, 0.4, 0.7 and 0.9 respectively, and predictors across different groups are independent. The true model includes two predictors from each group with coefficient 1, total $K = 8$ predictors in the model. The theoretical R^2 is set to 0.8. Similar to setting 2, we also consider three sample sizes: T =50, 100 and 200.

All simulations are conducted with 100 repetitions. For FS, IHT, HTP and SMC, the tuning parameter is the the subset size k, while for adaLASSO, the tuning parameter is the penalty factor λ. Both of them are selected by 5-fold cross validation. For the adaLASSO, We apply the R package `glmnet` and use LASSO as initial value to construct the penalty weight. For the GLSS+adaLASSO proposed by [63], at the screening stage, the proportion of predictors to be screened out is arbitrarily determined. Here we set two thirds of predictors to be screened out by GLSS, then adaLASSO is applied on the remaining one third predictors. For knockoff, literatures suggest FDR to be controlled within 0.1. However, we found knockoff will not select any predictor if we set FDR ≤ 0.1 in our simulation setting. Thus, we did a grid search of FDR from 0.2 to 0.7 and choose the one that gives the best MSPE and DC.

4.1.2. Simulation Results

Table 1 lists the five quantiles of the number of selected predictors by each VS method in setting 1, and the mean of R^2 in 100 repetitions. Recall that the true model contains 12 predictors by design. Clearly adaLASSO tends to pick too many predictors, with median of 38 and 34 under the two scenarios of theoretical R^2 respectively. Moreover, the range is much wider than those from other methods, reflecting its low stability in selecting predictors. This is not surprising since adaLASSO fails to distinguish between a zero and a nonzero coefficient when the corresponding predictors are correlated.

In contrast to adaLASSO, FS, SMC, IHT and HTP are more conservative and tends to select less predictors than the true number 12. For example, FS yields a median of 8 predictors and covers the range from 5 to 13. Such under-selection is actually expected, because cross validation is a conservative way to find the correct number of predictors by avoiding over-fitting. In addition, the under-selection does not ruin the model fitting performance that is measured by in-sample R^2. The R^2 obtained by FS, SMC, IHT and HTP are close to the theoretical R^2. To better understand the reason of under-selection, we calculate the hit ratio (the proportion of successfully identifying the predictor in 100 repetitions) of each predictor in the true model, as shown in Table 2. It appears that the predictors with small magnitudes of coefficient (0.1 and 0.4) would be excluded from the model due to their relatively low prediction power, especially when they are highly correlated with other predictors and the sample size is not very large ($T = 200$). It is also natural to see that the hit ratio is higher when the signal strength is higher (theoretical R^2=0.8) and the predictors are less correlated.

By controlling FDR or adding an initial variable screening, knockoff and GLSS+adaLASSO selected much less predictors compared to adaLASSO, which is expected. However, the consequence is the worse model fitting performance. Knockoff yields an inferior in-sample R^2 compared to other methods, as shown in Table 2. In particular, the minimum number of selected predictors is zero, which indicates no predictor is selected by knockoff in some simulation replicates. GLSS+adaLASSO is even much worse in terms of model fitting. In particular, it only yields in-sample $R^2 = 0.44$ when the theoretical R^2 is 0.8.

Table 2
Hit ratio of each predictor in the true model in setting 1. For each predictor, "corr" is the within-group correlation, "coef" is the coefficient, and hit ratio is the proportion of the 100 repetition that successfully identified the predictor.

corr	coef	FS	SMC	adaLASSO	IHT	HTP	knockoff	GLSS+adaLASSO
Theoretical $R^2 = 0.8$								
0.1	0.1	0.02	0.03	0.17	0.01	0.02	0.03	0.00
	0.4	0.44	0.38	0.82	0.15	0.42	0.37	0.00
	0.7	0.95	0.98	1.00	0.68	0.91	0.67	0.01
	1.0	1.00	1.00	1.00	0.97	0.98	0.77	0.07
0.4	0.1	0.01	0.01	0.14	0.01	0.03	0.01	0.00
	0.4	0.32	0.30	0.74	0.12	0.23	0.34	0.02
	0.7	0.89	0.93	0.99	0.60	0.70	0.57	0.22
	1.0	1.00	1.00	1.00	0.96	0.98	0.77	0.54
0.8	0.1	0.02	0.02	0.08	0.03	0.01	0.02	0.04
	0.4	0.06	0.04	0.27	0.10	0.02	0.13	0.26
	0.7	0.35	0.40	0.68	0.26	0.22	0.43	0.35
	1.0	0.79	0.88	0.97	0.51	0.45	0.63	0.75
Theoretical $R^2 = 0.5$								
0.1	0.1	0.01	0.02	0.09	0.01	0.01	0.04	0.00
	0.4	0.10	0.10	0.34	0.02	0.08	0.08	0.01
	0.7	0.41	0.42	0.79	0.25	0.36	0.32	0.03
	1.0	0.81	0.84	0.98	0.71	0.76	0.52	0.08
0.4	0.1	0.04	0.03	0.14	0.03	0.04	0.04	0.03
	0.4	0.13	0.18	0.35	0.10	0.10	0.10	0.11
	0.7	0.31	0.33	0.65	0.23	0.30	0.27	0.37
	1.0	0.72	0.73	0.92	0.63	0.69	0.50	0.56
0.8	0.1	0.01	0.00	0.03	0.01	0.00	0.00	0.01
	0.4	0.05	0.07	0.17	0.09	0.06	0.04	0.09
	0.7	0.08	0.11	0.28	0.17	0.07	0.10	0.18
	1.0	0.15	0.22	0.47	0.24	0.16	0.16	0.32

This is because the screening method, GLSS, looks at each predictor individually and marginally without considering correlation among them. To better explain this, we call the predictors in the true model true predictors, and call the predictors not in the true model null predictors. In our simulation setting, the 300 predictors in the third group (correlation = 0.8) are highly correlated with the four true predictors in that group. When looking at each predictor marginally and individually, each null predictor in that group has larger prediction power than the true predictors in the other two groups (correlation = 0.1 and 0.4). Thus, GLSS keeps these null predictors but screen out the true predictors in the first two groups. This can be clearly seen in Table 2: All the true predictors in the first two groups (correlation = 0.1 and 0.4) are much less selected by GLSS+adaLASSO, especially predictors in group one is seldom selected.

Table 3 summarizes the results of the five criteria in setting 1. When the signal is strong (theoretical $R^2 = 0.8$), SMC, adaLASSO and FS are the best group in terms of prediction performance, which are significantly superior to the other four methods. As for the variable selection accuracy, SMC and FS outperforms the rest. Since adaLASSO suffers from excessive over-selection, its precision and DC are low while its recall is the highest. HTP is a little better than IHT but still worse than SMC and FS. Because of the aforementioned screening issue in GLSS+adaLASSO, its prediction performance and variable selection accuracy is the worst among all the methods under investigation. As for time cost, SMC is the most expensive algorithm, which takes about three hours to finish one simulation while others cost less than one minute. FS and adaLASSO are the most time-efficient, which only take two seconds or less to finish one simulation. In summary, the most sophisticated algorithm, SMC, is the best in terms of both variable selection accuracy and out-of-sample prediction but extremely time consuming. FS, the classical greedy algorithm, obtains very similar performance as SMC while it is much more computationally efficient. If we only focus on prediction, adaLASSO is one of the best choices given its great efficiency and promising prediction results.

In the scenario of theoretical $R^2 = 0.5$, the differences in prediction among all five methods are small, while adaLASSO outperforms the rest a little bit. As for variable selection accuracy, GLSS+adaLASSO is still the worst, followed by knockoff and adaLASSO. On the other hands, SMC and FS are significantly superior to the rest and very close to each other. This scenario asserts that, when the signal is not strong enough, adaLASSO is preferred for prediction while FS and SMC can provide more accurate variable selection results.

In setting 2 and 3, we only report the results of MSPE and DC in Table 4 and 5, and leave other criteria in Appendix A (Table Appendix A.1 and Appendix A.2). It can be found that all the VS methods work better in the independent case (setting 2) than in dependent case (setting 3). In addition, as sample size T increases, their results become closer to that of the true model. In setting 2, all predictors are generated independently. So GLSS+adaLASSO does not suffer from aforementioned screening issue and behaves similarly to adaLASSO. In terms of sample size, $T = 50$ is too small to successfully recover the true sparsity or obtain a satisfying prediction for any VS method. When sample size increases to 100, SMC and FS have DC as 0.96 and 0.98 respectively, which means they almost completely recover the true model, and their prediction are also quite close to the true model. Knockoff is slightly falling behind, but significantly better than adaLASSO, IHT, HTP and

Table 3
Simulation results in Setting 1. the column "true" lists the results of the true model. For each criteria, the mean of 100 repetitions is presented, followed by the standard error in the parenthesis.

	true	FS	SMC	adaLASSO	IHT	HTP	knockoff	GLSS+adaLASSO
Theoretical $R^2 = 0.8$								
MSPE	2.43	3.07	2.97	2.99	3.44	3.55	4.99	5.89
	(0.04)	(0.05)	(0.05)	(0.05)	(0.08)	(0.09)	(0.32)	(0.14)
DC	–	0.59	0.61	0.32	0.40	0.48	0.38	0.20
	–	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
Precision	–	0.74	0.79	0.22	0.45	0.57	0.57[a]	0.23
	–	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.29)[a]	(0.01)
Recall	–	0.49	0.50	0.65	0.37	0.41	0.40	0.19
	–	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
time	–	0.04	168.25	0.01	0.29	0.75	12.7	8.12
	–	(0.00)	(0.76)	(0.00)	(0.00)	(0.00)	(1.29)	(0.24)
Theoretical $R^2 = 0.5$								
MSPE	9.59	12.58	12.79	11.33	11.96	12.66	14.02	12.95
	(0.16)	(0.24)	(0.24)	(0.17)	(0.20)	(0.24)	(0.35)	(0.22)
DC	–	0.30	0.31	0.23	0.25	0.27	0.18	0.14
	–	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
precision	–	0.41	0.41	0.16	0.32	0.37	0.34[a]	0.15
	–	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.03)[a]	(0.03)
recall	–	0.24	0.26	0.44	0.21	0.22	0.18	0.16
	–	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
time	–	0.04	162.68	0.02	0.29	0.75	15.9	8.14
	–	(0.00)	0.63	(0.00)	(0.00)	(0.00)	(4.0)	(0.23)

[a] In some simulation replicates, knockoff does not select any predictors. Precision of knockoff is calculated from those simulation replicates in which knockoff has selected at least one predictor.

Table 4
Out-of-sample MSPE and in-sample DC in setting 2. The column "true" lists the results of the true model. The mean of 100 repetitions is presented, followed by the standard error in the parenthesis.

T	true	FS	SMC	adaLASSO	IHT	HTP	knockoff	GLSS+adaLASSO
MSPE								
$T = 50$	1.35	5.96	5.43	4.78	5.69	5.84	5.83	5.1
	(0.03)	(0.20)	(0.22)	(0.14)	(0.13)	(0.14)	(0.14)	(0.15)
$T = 100$	1.32	1.41	1.36	2.26	2.87	2.56	1.58	2.46
	(0.02)	(0.03)	(0.03)	(0.06)	(0.11)	(0.12)	(0.08)	(0.07)
$T = 200$	1.28	1.29	1.29	1.60	1.33	1.28	1.36	1.75
	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.04)
DC								
$T = 50$	–	0.31	0.37	0.19	0.28	0.26	0.11	0.15
	–	(0.02)	(0.03)	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)
$T = 100$	–	0.96	0.98	0.19	0.70	0.76	0.88	0.15
	–	(0.01)	(0.00)	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)
$T = 200$	–	0.99	0.99	0.23	0.98	0.99	0.91	0.17
	–	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.01)	(0.01)

Table 5
Out-of-sample MSPE and in-sample DC in setting 3. The column "true" lists the results of the true model. The mean of 100 repetitions is presented, followed by the standard error in the parenthesis.

T	true	FS	SMC	adaLASSO	IHT	HTP	knockoff	GLSS+adaLASSO
MSPE								
$T = 50$	3.67	11.93	11.31	7.97	11.78	12.16	12.99	8.86
	(0.07)	(0.31)	(0.28)	(0.17)	(0.28)	(0.28)	(0.39)	(0.2)
$T = 100$	3.34	7.46	6.78	5.88	9.39	9.21	11.05	7.64
	(0.06)	(0.21)	(0.23)	(0.13)	(0.27)	(0.25)	(0.42)	(0.14)
$T = 200$	3.20	4.07	4.01	4.29	7.82	6.65	10.48	7.12
	(0.05)	(0.08)	(0.09)	(0.08)	(0.32)	(0.34)	(0.48)	(0.17)
DC								
$T = 50$	–	0.06	0.08	0.13	0.04	0.04	0.06	0.1
	–	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
$T = 100$	–	0.33	0.45	0.19	0.18	0.16	0.17	0.15
	–	0.02	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)
$T = 200$	–	0.69	0.74	0.24	0.29	0.46	0.18	0.25
	–	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.01)

Table 6
Settings of predictors

group i	ϕ_i	δ_{i1}^2	δ_{i2}^2	var(z_{ijt})	cor$(z_{ijt}, z_{ij't})$
1	0.7	0.357	0.3	1	0.7
2	0.7	0.153	0.7	1	0.3
3	0.3	0.637	0.3	1	0.7
4	0.3	0.273	0.7	1	0.3

GLSS+adaLASSO. When sample size increases to 200, adaLASSO and GLSS+adaLSSO still suffers from over-selection. Thus, their precision and DC are very low and MSPE are significantly worse than that of true model. Whereas, all other VS methods are quite close to the true model for both variable selection accuracy and out-of-sample prediction. In setting 3, since the potential predictors are correlated, all VS methods can not provide satisfying performance until sample size increases to 200. Under this sample size, FS and SMC have the smallest prediction error and provide the most accurate models, followed by adaLASSO whose prediction is close to SMC and FS but performs worse in variable selection. In both setting 2 and setting 3, when the sample size is not large enough ($T = 50$ in setting 2 and $T = 50$ or 100 in setting 3), adaLASSO has the best prediction, followed by SMC and GLSS+adaLASSO, and SMC is superior to others in terms of identifying the true predictors.

We summarize our findings in these three simulation studies as follows. (1) When the signal is strong and the sample size is large enough, SMC and FS are the best among the seven VS methods in terms of variable selection accuracy and prediction. However, SMC is very time consuming. The classical procedure FS is preferred if the time constraint is a concern. (2) adaLASSO is very good at prediction in all settings, especially when the signal is not strong or sample size is not large. However, it suffers from over-selection, and it is worse than others in terms of variable selection. (3) In general, IHT is slightly better than HTP, but both are significantly worse than SMC and FS. (4) When the sample size is too small or signal is not strong, none of these method can provide good performance, and their difference are insignificant. (5) knockoff and GLSS mitigate the overfitting issues in adaLASSO but compromise the model fitting and prediction performance. In particular, GLSS is not a good choice when predictors are correlated, since it does not consider correlation among predictors and screens predictors only based on their marginal associations with the target variable.

4.2. Comparison between VS and FA

In this simulation study, we investigate FA versus VS approaches under the framework of time series forecasting. The dimension of predictors and number of observations mimic the scales of our real data in Section 5. The simulation setting is described as follows.

We generate four groups of predictors with group id as $i = 1$, 2, 3 and 4. In each group, a series of latent factor $\{f_{it}\}$ is generated based on a AR(1) process, then 30 predictors are generated from this latent factor:

$$f_{it} = \phi_i f_{i,t-1} + e_{it}, \quad e_{it} \overset{\text{i.i.d}}{\sim} N\left(0, \delta_{i1}^2\right), \quad i = 1, \cdots, 4 \tag{31}$$

and

$$z_{ijt} = f_{it} + e_{ijt}, \quad e_{ijt} \overset{\text{i.i.d}}{\sim} N\left(0, \delta_{i2}^2\right), \quad j = 1, \ldots, 30, \quad t = 1, \ldots, T. \tag{32}$$

Clearly the series $\{z_{ijt}\}$ also follows a VAR(1) process with parameter ϕ_i. Within one group, the temporal dependence is captured by the AR parameter ϕ_i, while the cross-sectional correlation, cor$(z_{ijt}, z_{ij't})$, is controlled by the relative magnitude between var(f_{it}) and δ_{i2}^2. By setting the values of $(\phi_i, \delta_{i1}^2, \delta_{i2}^2)$ as shown in Table 6, the four groups are constructed to have different temporal and cross-sectional dependence. Specifically, both temporal and cross-sectional dependence are strong in group 1. While for group 2, only the temporal dependence is strong but the cross-sectional dependence is weak. Group 3 has the opposite pattern to group 2, and in group 4, both temporal and cross-sectional dependence are weak.

For the response variable, we consider two generating mechanisms:

$$\text{Setting 1: } y_t = \sum_{i=1}^{4}(0.8z_{i1t} + 0.4z_{i2t}) + \sum_{i=1}^{4}(0.8z_{i1,t-1} + 0.4z_{i2,t-1}) + \epsilon_t. \tag{33}$$

$$\text{Setting 2: } y_t = \sum_{i=1}^{4} 0.8 f_{it} + \sum_{i=1}^{4} 0.4 f_{i,t-1} + \varepsilon_t. \tag{34}$$

Setting 1 indicates that y_t is a sparse linear function of the predictors and their lagged values, which favors VS approach over FA. In contrast, setting 2 implies that y_t is directly generated from the latent factors, thus FA is the correct approach while VS approach miss-specifies the model. In both settings, the theoretical R^2 is set to be 0.8. In each setting, we generate 300 observations ($T = 300$).

Table 7
Results of FA and VS approaches

	true	FA_BIC	FA_FCV	FS	SMC	adaLASSO	IHT	knockoff	GLSS+adaLASSO
Setting 1: y_t generated from predictors									
MSPE	3.08	6.52	6.76	4.25	4.16	3.96	5.76	9.07	4
	(0.07)	(0.14)	(0.15)	(0.11)	(0.10)	(0.08)	(0.23)	(0.48)	(0.1)
R^2	0.81	0.61	0.63	0.80	0.82	0.84	0.72	0.36	0.85
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.04)	(0.00)
DC	–	–	–	0.67	0.71	0.50	0.50	0.29	0.42
	–	–	–	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)	(0.01)
Precision	–	–	–	0.78	0.77	0.36	0.55	0.73[a]	0.29
	–	–	–	(0.02)	(0.02)	(0.01)	(0.01)	(0.19)[a]	(0.01)
Recall	–	–	–	0.61	0.68	0.87	0.48	0.29	0.84
	–	–	–	(0.01)	(0.01)	(0.01)	(0.02)	(0.03)	(0.01)
Setting 2: y_t generated from factors									
MSPE	0.57	0.68	0.71	1.20	1.14	0.82	1.27	2.48	0.87
	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.09)	(0.02)
R^2	0.81	0.79	0.80	0.80	0.81	0.88	0.71	0.12	0.88
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.02)	(0.00)

[a] In some simulation replicates, knockoff does not select any predictors. Precision of knockoff is calculated from those simulation replicates in which knockoff has selected at least one predictor.

We select FS, SMC, adaLASSO, IHT, knockoff and GLSS+adaLASSO as representatives of VS approach and compare them with FA. For both FA and VS approaches, we do not include the lagged value of y_t as predictor. Thus the models for FA and VS approaches are:

$$\text{FA approach:} \quad y_t = \sum_{l=0}^{m-1} \boldsymbol{\gamma}_l' \boldsymbol{f}_{t-l} + \varepsilon_t, \quad \boldsymbol{Z}_t = \boldsymbol{\Lambda} \boldsymbol{F}_t + \boldsymbol{e}_t. \tag{35}$$

$$\text{VS approach:} \quad y_t = \boldsymbol{X}_t' \boldsymbol{\beta} + \epsilon_t. \tag{36}$$

To predict y_t, the vector $\boldsymbol{X}_t$ in VS approach consists of z_{ijt}'s and their historical values up to lag 5, i.e. $\boldsymbol{X}_t = \{z_{11t}, z_{11,t-1}, \ldots, z_{11,t-5}, \ldots, z_{4,30,t}, z_{4,30,t-1}, \ \ldots, z_{4,30,t-5}\}$, with total of 720 predictors. While the vector $\boldsymbol{Z}_t$ in FA approach only contains z_{ijt}'s, i.e. $\boldsymbol{Z}_t = \{z_{11t}, \ldots, z_{4,30,t}\}$. The tuning parameters in VS methods are selected by forward cross-validation (FCV). For FA approach, we consider both BIC and FCV to determine the number of factors d and the order of lagged factors m. The corresponding results are labeled as FA_BIC and FA_FCV respectively.

FCV is commonly used for tuning parameter selection in time series analysis [52,48,62], which selects parameters by simulating real-time out-of-sample forecasting. Take adaLASSO for example. First, the last 50 observations ($t = 251, \ldots, 300$) are preserved as test set for evaluating out-of-sample forecasting. Then we divide the left data ($t = 1, \ldots, T_1 = 250$) into two sets: training set ($t = 1, \ldots, T_0$) and validation set ($t = T_0 + 1, \ldots, T_1$). In this simulation section, we set $T_0 = 200$. We follow [30] to perform the grid search of λ. For each value of λ, we obtains the adaLASSO estimation from the training set and then carry out prediction on the validation set. The optimal tuning parameter λ_{opt} is selected by minimizing the following prediction errors on the validation set:

$$\frac{1}{T_1 - T_0} \sum_{t=T_0+1}^{T_1} (y_t - \hat{y}_t^{\lambda})^2, \tag{37}$$

where $\hat{y}^{\lambda}$ is the forecasting value of y_t. After that, the adaLASSO is refit using λ^{opt} and data till T_1 to obtain the final model estimation. At last, we use this final estimation to make forecast on the test set $\{y_t : t = T_1 + 1, \ldots, 300\}$ and the corresponding forecasting error is used to calculate the criteria MSPE.

Table 7 shows the results of comparison between FA and VS approaches. Unsurprisingly, VS methods are uniformly better than FA in setting 1 and are inferior to FA in setting 2. For FA approach, there is little difference between the two criteria, BIC and FCV. While among the six VS methods, adaLASSO and GLSS+adaLASSO yield the most precise prediction followed by SMC and FS with no significant difference. As for variable selection accuracy, SMC and FS are the best while adaLASSO over-selects predictors. Based on this simulation study, we can conclude that, if only a handful of predictors are relevant to the response variable, VS methods are more advantageous than FA approach. However if there are many relevant predictors which possess a factor structure, FA approach is more suitable.

5. Real-Time Macroeconomic Forecasting

In this section, we apply both FA and VS approaches on forecasting of several important macroeconomic indices. By simulating real-time forecasting, we compare their empirical performance in real application. In addition, intersection between

machine learning (ML) and economic forecasting becomes an important and attractive research area [31,21,44]. In this paper, we choose random forest (RF) as an representative of ML techniques and compare its forecasting performance with the aforementioned FA and VS approaches. RF is an ensemble ("bagging") machine learning approach which constructs a multitude of regression trees by bootstrap and then take average of them. Because of this bagging feature and its capability of handling nonlinearity and interaction among predictors, RF performs well in many prediction problems. However, the fitted RF model is an aggregation of hundreds of regression trees and each tree is a nonparametric model. Thus, it is much less interpretable than the parametric model constructed by FA and VS approaches.

Our target indices are employment (EMP), industrial production index (IP) and consumer price index-all urban consumers (CPI). A well-known large-scale macroeconomic database, called FRED-MD, is used as the pool of predictors. FRED-MD database is constructed by [42], designed as a comprehensive and well-maintained database in order to facilitate "big data" analysis in macroeconomic research. It includes 134 monthly economic variables and covers main aggregates, coincident indicators and leading economic indicators, which mimic the coverage of datasets already in the literature [11,54,53]. Additionally, it has three appealing features: (1) monthly updated in real-time; (2) publicly accessible in research.stlouisfed.org/econ/mccracken/fred-databases ; (3) facilitating the replication of empirical works.

5.1. Implementation Details

We follow the framework of economic forecasting stated in Section 2. Let y_{t+h}^h be the h-month ahead value of the variable to be forecasted and y_t be the corresponding t-dated value, which are required to be stationary. Following [55] and [3], we define y_{t+h}^h and y_t as follows:

$$y_{t+h}^h = (1200/h)ln(EMP_{t+h}/EMP_t),\ y_t = 1200ln(EMP_t/EMP_{t-1}), \tag{38}$$

$$y_{t+h}^h = (1200/h)ln(IP_{t+h}/IP_t),\ y_t = 1200ln(IP_t/IP_{t-1}), \tag{39}$$

$$\begin{aligned} y_{t+h}^h &= (1200/h)ln(CPI_{t+h}/CPI_t) - 1200ln(CPI_t/CPI_{t-1}), \\ y_t &= 1200 \triangle ln(CPI_t/CPI_{t-1}). \end{aligned} \tag{40}$$

5.1.1. Data Processing

We exclude 6 variables from FRED-MD which have quite a few missing data during our study period (1994:1 to 2018:12). For the remaining 128 variables, we carry out the transformation proposed by [42] to make them stationary. Then each variable is standardized to have mean zero and variance one.

After the transformation and standardization, the 128 economic variables compose the input vector $\boldsymbol{Z}_t$ of factor model (2) in the FA approach. Following existing empirical works [54,55,3,42], the prediction model (3) includes the lagged value of both the target index y and the factors $\boldsymbol{f}_t$ up to half a year, i.e. $q \leq 6$ and $m \leq 6$. To gain a fair comparison, for the VS methods, the potential predictors also contain the lag value of $\boldsymbol{Z}_t$ up to half a year, i.e. $\boldsymbol{X}_t = (\boldsymbol{Z}_t, \ldots, \boldsymbol{Z}_{t-5})$, with dimension 768. Since both FA and VS approaches are under the framework of linear regression, their coefficient estimation is sensitive to outliers. According to [54,55], in a series $\{z_{it}\}$, an outlier is defined as an observation that deviates from the sample median by more than ten interquantile ranges. After identifying the outliers, FA approach replaces them with missing values (NA), and estimates the factors using EM algorithms to account for missing values. As for VS approach, we can either delete the series contaminated by outliers or impute the values by kalman filter. Therefore, we consider two strategies to account for outliers in the predictors:

1. Remove the series with outliers: For fair comparison, we also delete these series in both VS and FA approaches;
2. For FA approach, replace outliers with missing values and implement the EM algorithm. For VS methods, impute the outliers using kalman filter in R package `imputeTS`.

For the target variable y, it does not contain any outlier, i.e. no value is beyond the ten interquantile ranges from the median. In our empirical results, generally FA and VS approaches perform slightly better under strategy 1. Thus we only report the results of strategy 1 in the main context, and put the results of strategy 2 in Table Appendix A.4.

5.1.2. Tuning Parameter Selection

There are three tuning parameters in FA approach: number of factors d, lags of y_t, q, and lags of the factors m. We choose (d, p, m) from $\{0 \leq d \leq 5, 0 \leq q \leq 6, 1 \leq m \leq 6\}$ by FCV, where $d = 0$ means no factor is included in the forecasting equation (3) and $p = 0$ means no auto-regressive term is included. For VS methods, the tuning parameter for FS, SMC, IHT and HTP is the number of selected predictors, and the tuning parameter for adaLASSO and GLSS+adaLASSO is the penalty λ. Both of them are determined by FCV. For RF, we use the default setting in `randomForest` package in R: The number of bootstrap samples is 500; each individual tree in the RF model is grown until there are only five observations in each

Table 8
Ratio between MSPE of different approaches and MSPE of the benchmark model (41). For each horizon h, the top three ratios are labeled in bold.

	FA	FS	SMC	adaLASSO	IHT	HTP	knockoff	GLSS+adaLASSO	RF
EMP									
h=1	0.95	**0.90**	0.94	0.94	0.97	0.97	0.91	**0.90**	**0.80**
h=3	**0.68**	0.74	**0.66**	0.79	0.72	**0.70**	0.82	0.73	0.73
h=6	0.88	**0.68**	**0.68**	0.76	0.71	**0.67**	0.84	0.78	**0.44**
h=12	0.62	0.69	**0.30**	**0.29**	0.85	0.78	0.84	0.32	**0.20**
IP									
h=1	0.99	**0.81**	**0.78**	0.91	**0.83**	**0.83**	0.92	0.90	0.88
h=3	**0.67**	**0.63**	0.85	0.84	0.83	0.83	1.12	**0.82**	0.89
h=6	0.93	**0.87**	**0.89**	1.24	**0.81**	0.90	1.72	1.26	1.11
h=12	**1.00**	1.27	1.72	**1.13**	1.58	1.56	1.44	**1.13**	**0.69**
CPI									
h=1	**0.95**	1.13	1.10	1.04	**0.96**	**0.98**	1.06	1.16	1.17
h=3	0.98	0.88	1.09	**0.78**	**0.75**	**0.76**	0.97	0.89	0.97
h=6	1.05	**0.82**	0.91	**0.84**	**0.77**	1.00	1.26	**0.84**	1.37
h=12	1.26	1.12	**0.81**	1.18	**0.81**	1.14	**0.95**	1.21	1.72

leaf; the proportion of variables randomly selected in each split is 1/3. In addition, we include univariate AR as benchmark model in the comparison:

$$y_{t+h}^h = \alpha + \sum_{l=0}^{\tilde{p}-1} \alpha_l y_{t-l} + \varepsilon_{t+h}, \tag{41}$$

where the AR order $\tilde{p}$ is selected from $0 \le \tilde{p} \le 6$ by FCV.

5.1.3. Simulating Real-Time Forecasting

The time span from 2015:1 to 2018:12 is reserved as test set to evaluate out-of-sample forecasting, with forecast horizon h being 1, 3, 6 or 12 months. For both FA and VS approaches, the estimations and forecasts are conducted to simulate real-time forecasting using a rolling window scheme with window size 240. For example, if we want to forecast the h-month growth rate of EMP on 2015:1, The target variable is y_{t+h}^h with $t = 2015{:}1 - h$. Here $2015{:}1 - h$ stands for the month that is h months before 2015:1. For example, if $h = 3$, $2015{:}1 - h$ is 2014:10. To obtain its forecast, both FA and VS use data from $1995{:}1 - h$ to $2015{:}1 - h$ (240 months) to estimate model and select tuning parameters. The validation set for selecting tuning parameters is the sub-sample from $2011{:}1 - h$ to $2015{:}1 - h$ (48 months). When forecasting y_{t+h}^h on 2015:2 ($t = 2015{:}2 - h$), the tuning parameters are re-selected and the models are re-estimated using data from $1995{:}2 - h$ to $2015{:}2 - h$ (validation set is $2011{:}2 - h$ to $2015{:}2 - h$). In this way, we allow the parameter estimation and the optimal number of predictors and factors to change across t. For each pair of target variable and forecasting horizon h, a method produces 48 forecasts on the test set. We use MSPE of these 48 forecasts to evaluate the forecasting performance of this method. The ratio between MSPE of each method and that of the benchmark model (41) is reported. Ratio less than one indicates the method has smaller forecasting error than the univariate AR.

5.2. Empirical Results

The MSPE ratios are presented in Table 8. For each forecast horizon h, we define the methods with the three smallest MSPE ratios as the best group. They are marked in bold. For each target variable, We only report the predictors selected with at least 12 times in the 48 rolling-window forecasts. The predictors with frequency less than 12 are viewed as lacking of systematic prediction power. The predictors selected by the best VS method and their number of occurrences in the 48 forecasts are reported in Table Appendix A.3.

For EMP, all the ratios are smaller than one, which manifests the usefulness of incorporating many predictors in forecasting. Among all the approaches, RF yields good forecasting performance and ranks the 1st in 1, 6 and 12-months ahead forecasting. Within VS methods, SMC is the best overall and outperforms FA for all the forecasting horizons. For $h = 1$ and 3, the improvement of SMC over FA is not obvious, while for $h = 6$ and 12, SMC reduces the MSPE of FA approach by 33% and 52% respectively.

The success of SMC implies EMP can be forecasted by only a few predictors. This point is also demonstrated in Table Appendix A.3, which shows that only two to five predictors are selected for each horizon. Specifically, t-dated EMP is selected in almost all 48 forecasts for all horizons, which implies the historical value has significant prediction power for EMP. For $h = 6$ and 12, the linear forecasting model also includes two interest rates: 3-month treasury C minus FEDFUNDS and 3-month commercial paper minus FEDFUNDS. The leading effect of interest rate on EMP can be explained by economic theory. With low interest rate, consumers are more likely to consume now rather than wait for later. Low interest rate also drop the cost of borrowing to invest. Thus the increase in consumption and investment leads to higher demand for labor. As for one-month ahead forecast ($h = 1$), except for the t-dated value of EMP, the model also includes M2 money stock, real personal consumption expenditures and two variables related to IP (IP:Fuel and IP:final products and nonindustrial supplies).

In terms of IP, neither FA nor VS approaches improve one-year-ahead forecasting ($h = 12$) over AR, while RF outperforms AR. A possible explanation could be that RF can account for nonlinearity and interaction among predictors. While for short-term forecasts ($h = 1$, 3 and 6), both FA and VS methods are better than AR with the VS method FS performing either the best or close to the best among all the methods. In addition, FS surpass FA by reducing MSPE by 18%, 6% and 6% for 1, 3 and 6-months ahead forecast respectively. Among the predictors selected by FS, IP:durable subcategory, S&P's composite common stock:dividend yield and 3-month treasury C minus FEDFUNDS occurs the most often. Some other subcategories of IP are also frequently selected (more than 24 times). This leads us to conclude that, besides its own subcategories, the stock market and interest rate can forecast the movement of IP.

The results of CPI show that, across all horizons, FA does not have significant advantage over AR, and even worse than it for one-year ahead forecast. This finding is consistent with the results in [42]. Furthermore, RF is also inferior to AR, especially for 6 and 12-months ahead forecast. In contrast, some of the VS methods do achieve substantial improvement over AR, especially the IHT algorithm. IHT performs the best among all the methods across all horizons, which reduces the MSPE of AR by 25%, 23% and 19% for 3, 6 and 12-months ahead forecasts respectively. Comparing IHT with FA, the improvement is even more apparent: IHT reduce the MSPE of FA by 23%, 27% and 36% for 3, 6, and 12-months ahead forecasts respectively.

The predictors selected by IHT are listed in Table Appendix A.3. For short-term forecast ($h = 1$, 3 and 6), only a handful of predictors are included in the model, and the predictors occurred most are real M2 money stock, t-dated CPI and its transportation subcategory, and oil price. This finding is consistent with economic theory and the composition of CPI. First, the money supply M2 has prediction power for CPI. This argument is supported by the classical quantity theory of money [45] and evidenced by some empirical studies [2,3]. Second, since CPI series has temporal dependence, it is natural to use current value (t-dated CPI) to forecast its future value. Third, transportation is the second largest category in CPI and is very sensitive to the oil price. Oil price also has direct impact on prices of many industrial materials which are the upstream prices of consumer price. In addition, oil price has great influence on other aspects of US economy such as stock market and investment. All these points make oil price an important index for CPI forecasting.

To summarize, there is no one approach uniformly outperforming the others. The best four approaches are RF, FS, SMC and IHT, since they are more frequently selected into the best group than others. Compared with FA approach, several VS methods yield better forecasts, especially in the forecasting of EMP and CPI where SMC and IHT improve upon FA to a large extent. Moreover, for each target variable, the predictors selected by the best VS method are consistent with the underlying economic theory, which highlights the good interpretability of VS approach. RF can gain improvement over AR and FA approaches in forecasting of EMP, while the fitted RF model is an aggregation of 500 tree models thus is not interpretable.

6. Conclusion and Discussion

FA and VS approach indicate two different directions in economic forecasting. FA approach implies the target variable has many relevant predictors which can be explained by a few latent factors, while VS approach assumes a handful of predictors have prediction powers on the target variable. Which approach is the best depends on the true data structure and the target variable to be forecasted. This paper aims to draw readers' attention on VS approach, which is less emphasized in the economic literature. In this paper, we introduce several cutting-edge VS algorithms to economic forecasting and compare to FA approach. It turns out for some target variables, VS approach is superior to FA approach. In particular, SMC significantly outperforms FA in forecasting of EMP, FS is superior to FA and AR for short-term forecasting of IP, and IHT is the best when predicting CPI. These methods also provide interpretable models which well explain the relationship between the target variable and the selected predictors. The second contribution of this paper lies on the overview and comparison among five different groups of VS methods. The last two groups, GDS-type algorithms and meta-heuristic algorithms, are popular in computer science but have not been widely used in economic forecasting. Several simulation studies are conducted to compare their prediction performance and variable selection accuracy. The most interesting finding is that, in all the simulation studies, the classical procedure FS works pretty well and sometimes even better than some advanced algorithms. Among all the reviewed VS methods, only the very time-consuming algorithm, SMC, is slightly but not significantly better than FS.

In economic forecasting, it is often the case that the underlying data structure is very complex. The relationships among different economic variables may vary in different time period. Therefore, it is unrealistic to expect any approach to be uniformly the best. It is an interesting and important research topic that how to practically choose the best prediction model. In real-time economic forecasting, if an interpretable model is preferred, we recommend to try FS, SMC, IHT based on our observation in simulation studies and real data application. After that, model selection or model averaging can be applied on these three models to construct the final model. For example, if we want to forecast the value of CPI on 2022:01 when we are on 2021:12 (i.e. 1-month ahead forecast), we can use a close period (e.g. 2019:01 to 2021:12) as test data set to simulate 1-month ahead forecast and select the best model according to the MSPE. If interpretable model is not required, machine learning algorithms such as RF and Xgboost can be included as well.

There has always been a debate that if VS approach is better than FA approach or not. [32] investigate a Bayesian predictive modeling using "spike-and-slab" prior and apply it on six popular datasets. Based on the posterior distributions, they find it is not usually possible to identify sparse predictive models by selecting a handful of predictors. Whereas, [41] show LASSO-type approaches outperform FA approach on forecasting of 20 macroeconomic variables. Even though VS approaches outperform FA approach in our empirical studies, it does not mean VS approach is always the right direction. For a target

variable, it could be the case that the nonzero predictors are combination of a handful of strong predictors (with large coefficients) and many weak predictors (with small coefficients). When the contribution of the weak predictors is small enough, a sparse model would be a good approximation and provide satisfying prediction, while a dense model will introduce too much noise thus is inferior. If the weak contributors are not negligible, FA could be a better solution than VS approach. Thus, it is appealing to combine the ideas of both VS and FA approaches into one method. In the past two decades, several works have been developed on this path, called supervised factor models (SFM), which includes targeted predictor approach [3], supervised principle component analysis [4] and combining forecasts using principal components [59]. These methods take into consideration the target variable when estimating the latent factors. For example, [3] and [59] first apply least angle regression to select relevant predictors and then construct factors only within the selected predictors. Most of existing methods only consider the commonly used l_1 regularization method, and little attention has been paid on l_0 constraint optimization algorithms such as the IHT and SMC algorithms. SFM using l_0 constraint optimization is an approach worthy for investigation.

Declaration of Competing Interest

No potential conflict of interest is declared by the authors.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecosta.2023.01.003.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6), 716–723.
Bachmeier, L.J., Swanson, N.R., 2005. Predicting inflation: Does the quantity theory help? Economic Inquiry 43 (3), 570–585.
Bai, J., Ng, S., 2008. Forecasting economic time series using targeted predictors. Journal of Econometrics 146 (2), 304–317.
Bair, E., Hastie, T., Paul, D., Tibshirani, R., 2006. Prediction by supervised principal components. Journal of the American Statistical Association 101 (473), 119–137.
Barber, R.F., Candès, E.J., 2015. Controlling the false discovery rate via knockoffs. The Annals of Statistics 43 (5), 2055–2085.
Barber, R.F., Candès, E.J., 2019. A knockoff filter for high-dimensional selective inference. The Annals of Statistics 47 (5), 2504–2537.
Basu, S., Michailidis, G., 2015. Regularized estimation in sparse high-dimensional time series models. The Annals of Statistics 43 (4), 1535–1567.
Bates, S., Candès, E., Janson, L., Wang, W., 2021. Metropolized knockoff sampling. Journal of the American Statistical Association 116 (535), 1413–1427.
Bańbura, M., Giannone, D., Modugno, M., Reichlin, L., 2013. Now-casting and the real-time data flow. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting. In: Handbook of Economic Forecasting, Vol. 2. Elsevier, pp. 195–237.
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) 57 (1), 289–300.
Bernanke, B.S., Boivin, J., 2003. Monetary policy in a data-rich environment. Journal of monetary economics 50 (3), 525–546.
Bertsimas, D., King, A., Mazumder, R., 2016. Best subset selection via a modern optimization lens. The annals of statistics 44 (2), 813–852.
Bertsimas, D., King, A., Mazumder, R., 2016. Best subset selection via a modern optimization lens. The Annals of Statistics 44 (2), 813–852.
Blumensath, T., Davies, M.E., 2009. Iterative hard thresholding for compressed sensing. Applied and Computational Harmonic Analysis 27 (3), 265–274.
Brooks, S.P., Friel, N., King, R., 2003. Classical model selection via simulated annealing. Journal of the royal statistical society: Series b (statistical methodology) 65 (2), 503–520.
Brusco, M.J., 2014. A comparison of simulated annealing algorithms for variable selection in principal component analysis and discriminant analysis. Computational Statistics & Data Analysis 77, 38–53.
Bühlmann, P., van de Geer, S., 2011. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Berlin Heidelberg.
Candes, E., Fan, Y., Janson, L., Lv, J., 2018. Panning for gold:-model-x-knockoffs for high dimensional controlled variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80 (3), 551–577.
Cerny, V., 1985. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. Journal of Optimization Theory and Applications 45 (1), 41–51.
Chatterjee, S., Laudato, M., Lynch, L.A., 1996. Genetic algorithms and their statistical applications: an introduction. Computational Statistics & Data Analysis 22 (6), 633–651.
Coulombe, P.G., Leroux, M., Stevanovic, D., Surprenant, S., 2019. How is machine learning useful for macroeconomic forecasting? Journal of Applied Econometrics.
Dai, W., Milenkovic, O., 2009. Subspace pursuit for compressive sensing signal reconstruction. IEEE Transactions on Information Theory 55 (5), 2230–2249.
De Mol, C., Giannone, D., Reichlin, L., 2008. Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? Journal of Econometrics 146 (2), 318–328. doi:10.1016/j.jeconom.2008.08.011. https://www.sciencedirect.com/science/article/pii/S0304407608001103
Duan, J.-C., 2019. Variable selection with big data based on zero norm and via sequential monte carlo. Available at SSRN: https://ssrn.com/abstract=3377038 or https://doi.org/10.2139/ssrn.3377038.
Duan, J.-C., Zhang, C., 2015. Non-gaussian bridge sampling with an application. Available at SSRN 2675877.
ECB, 2008. Short-term forecasts of economic activity in the euro area. Working Paper. European Central Bank.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96 (456), 1348–1360.
Foucart, S., 2011. Hard thresholding pursuit: An algorithm for compressive sensing. SIAM Journal on Numerical Analysis 49 (6), 2543–2563.
Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., 2007. Pathwise coordinate optimization. Annals of Applied Statistics 1 (2), 302–332. doi:10.1214/07-AOAS131.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33 (1), 1.
Garcia, M.G.P., Medeiros, M.C., Vasconcelos, G.F.R., 2017. Real-time inflation forecasting with high-dimensional models: The case of brazil. International Journal of Forecasting 33 (3), 679–693.
Giannone, D., Lenza, M., Primiceri, G.E., 2021. Economic predictions with big data: The illusion of sparsity. Econometrica 89 (5), 2409–2437.
Giannone, D., Reichlin, L., Banbura, M., Modugno, M., 2013. Now-casting and the real-time data flow. Working Paper Series. European Central Bank.

Giannone, D., Reichlin, L., Small, D., 2008. Nowcasting: The real-time informational content of macroeconomic data. Journal of Monetary Economics 55 (4), 665–676.
Goldberg, D.E., 2006. Genetic algorithms. Pearson Education India.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer New York.
Jacob, L., Obozinski, G., Vert, J.-P., 2009. Group lasso with overlap and graph lasso. In: Proceedings of the 26th annual international conference on machine learning, pp. 433–440.
Jain, P., Tewari, A., Dhillon, I.S., 2011. Orthogonal matching pursuit with replacement. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 24. Curran Associates, Inc., pp. 1215–1223.
Jain, P., Tewari, A., Kar, P., 2014. On iterative hard thresholding methods for high-dimensional m-estimation. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 685–693.
Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220 (4598), 671–680.
Li, J., Chen, W., 2014. Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models. International Journal of Forecasting 30 (4), 996–1015.
McCracken, M.W., Ng, S., 2016. FRED-MD: A monthly database for macroeconomic research. Journal of Business & Economic Statistics 34 (4), 574–589.
McCracken, M.W., Ng, S., 2020. FRED-QD: A Quarterly Database for Macroeconomic Research. Working Paper. Federal Reserve Bank of St. Louis.
Medeiros, M.C., Vasconcelos, G.F.R., Veiga, A., Zilberman, E., 2021. Forecasting inflation in a data-rich environment: the benefits of machine learning methods. Journal of Business & Economic Statistics 39 (1), 98–119.
Mill, J.S., 1965. Principles of political economy.
Needell, D., Tropp, J.A., 2009. CosaMP: Iterative signal recovery from incomplete and inaccurate samples. Applied and Computational Harmonic Analysis 26 (3), 301–321.
Ng, S., 2013. Chapter 14 - variable selection in predictive regressions. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting. In: Handbook of Economic Forecasting, Vol. 2. Elsevier, pp. 752–789.
Nicholson, W.B., Wilms, I., Bien, J., Matteson, D.S., 2020. High dimensional forecasting via interpretable vector autoregression. J. Mach. Learn. Res. 21 (166), 1–52.
Puig, A.T., Wiesel, A., Hero, A.O., 2009. A multidimensional shrinkage-thresholding operator. In: 2009 IEEE/SP 15th Workshop on Statistical Signal Processing. IEEE, pp. 113–116.
Romano, Y., Sesia, M., Candès, E., 2020. Deep knockoffs. Journal of the American Statistical Association 115 (532), 1861–1872.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464. doi:10.1214/aos/1176344136.
Song, S., Bickel, P.J., 2011. Large vector auto regressions. arXiv preprint arXiv:1106.3915.
Stock, J.H., Watson, M., 2011. Dynamic factor models. Oxford Handbooks Online.
Stock, J.H., Watson, M.W., 2002. Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97 (460), 1167–1179.
Stock, J.H., Watson, M.W., 2002. Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics 20 (2), 147–162.
Stock, J.H., Watson, M.W., 2006. Chapter 10 forecasting with many predictors. In: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting. In: Handbook of Economic Forecasting, Vol. 1. Elsevier, pp. 515–554.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58 (1), 267–288.
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (1), 91–108.
Tu, Y., Lee, T.-H., 2019. Forecasting using supervised factor models. Journal of Management Science and Engineering 4 (1), 12–27.
Wainwright, M.J., 2019. High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
Wang, H., Li, G., Tsai, C.-L., 2007. Regression coefficient and autoregressive order shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 (1), 63–78.
Wang, Z., Safikhani, A., Zhu, Z., Matteson, D.S., 2020. Regularized estimation in high-dimensional vector auto-regressive models using spatio-temporal information. arXiv preprint arXiv:2012.10030.
Yousuf, K., 2018. Variable screening for high dimensional time series. Electronic Journal of Statistics 12 (1), 667–702.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (1), 49–67.
Zhang, C.-H., 2010. Nearly unbiased variable selection under minimax concave penalty. The Annals of statistics 38 (2), 894–942.
Zhao, P., Rocha, G., Yu, B., 2008. The composite absolute penalties for grouped and hierarchichal variable selection. Annals of Statistics.(to appear).
Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101 (476), 1418–1429.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 67 (2), 301–320.