axioms

MDPI

Article

Analysis of Fat Big Data Using Factor Models and Penalization Techniques: A Monte Carlo Simulation and Application

Faridoon Khan [1,*] and Olayan Albalawi [2]

[1] Department of Creative Technology, Faculty of Computing and AI, Air University, Islamabad 44000, Pakistan
[2] Department of Statistics, Faculty of Science, University of Tabuk, Tabuk 47512, Saudi Arabia; oalbalwi@ut.edu.sa
* Correspondence: faridoon.khan@au.edu.pk

Abstract: This article assesses the predictive accuracy of factor models utilizing Partial·Least·Squares (PLS) and Principal·Component·Analysis (PCA) in comparison to autometrics and penalization techniques. The simulation exercise examines three types of scenarios by introducing the issues of multicollinearity, heteroscedasticity, and autocorrelation. The number of predictors and sample size are adjusted to observe the effects. The accuracy of the models is evaluated by calculating the Root·Mean·Square·Error (RMSE) and the Mean·Absolute·Error (MAE). In the presence of severe multicollinearity, the factor approach utilizing (PLS demonstrates exceptional performance in comparison. Autometrics achieves the lowest RMSE and MAE values across all levels of heteroscedasticity. Autometrics provides better forecasts with low and moderate autocorrelation. However, Elastic·Smoothly·Clipped·Absolute·Deviation (E-SCAD) forecasts well with severe autocorrelation. In addition to the simulation, we employ a popular Pakistani macroeconomic dataset for empirical research. The dataset contains 79 monthly variables from January 2013 to December 2020. The competing approaches perform differently compared to the simulation datasets, although "The PLS factor approach outperforms its competing approaches in forecasting, with lower RMSE and MAE". It is more probable that the actual dataset exhibits a high degree of multicollinearity.

Keywords: fat big data; factor models; machine learning techniques; forecasting; Monte Carlo experiments; inflation

MSC: 94D05

Citation: Khan, F.; Albalawi, O. Analysis of Fat Big Data Using Factor Models and Penalization Techniques: A Monte Carlo Simulation and Application. *Axioms* **2024**, *13*, 418. https://doi.org/10.3390/axioms13070418

Academic Editors: Tamer M Elbayoumi and Tae Yoon Kim

Received: 7 May 2024
Revised: 18 June 2024
Accepted: 18 June 2024
Published: 21 June 2024

1. Introduction

Regression analysis is a widely recognized statistical method employed in various fields, including finance and the social sciences. The main objective of regression analysis is to create a model that accurately represents the influence of one or more independent variables on a dependent variable. The Ordinary·Least·Squares (OLS) approach is a frequently employed technique for estimating unknown parameters of a regression model [1]. The OLS estimates are derived by reducing the squared errors of the residuals. The approach is widely favored because of its high interpretability and ability to generate accurate estimates, provided that the underlying assumptions are met [2].

In the era of big data, dataset formats have changed. Previously, the number of observations, n, was generally much bigger than the number of explanatory variables, p. However, currently, $n \approx p$ or even $n < p$ is common, referred to as high-dimensional data. These large datasets have presented new issues, such as degrees of freedom, multicollinearity, heteroscedasticity, etc., rendering standard linear regression models ineffective. Traditional econometric models do not provide sparse models, which may result in inefficient behavior when $n < p$. Advanced regression approaches are consequently necessary for enormous datasets, commonly known as big data [3].

The recent developments in the collection of macroeconomic data have led to a great focus on big data. An accurate analysis can be performed if we extract the important information suitably from a huge set of features. However, the performance alters depending on the data dimension and estimation tool which are applied as well. Failure in dimensional reduction induces poor output because of redundant variables. Since influential work on the forecasting through Diffusion·Index (DI) was conducted by [4], factors models are considered the common approach for predictive modeling in a data-rich environment. Stock and Watson [5] showed that forecasting via factor models is more accurate than the existing forecasting tools like autoregressive forecasts, bagging, pretest methods, empirical Bayes, and Bayesian model averaging. They inferred that the DI is an effective approach to lessen the regression dimension, and it appears to be difficult to enhance this performance without introducing severe changes to the predictive model. Recently, the factor models extended for forecasting aims include those of [6–10].

In addition to the DI methodology, sparse regression is another family of tools utilized for dimension reduction and forecasting, and it is specifically well-known in the econometrics and statistics fields. The sparse regression tools attempt to keep the relevant features and force the coefficients of irrelevant features to zero. The benefit of such tools is that they permit a curse of dimensionality that is available in macroeconomic time series for a substantial amount of time, but the predictions that statistical tools produce also serve to devise productive monetary policies [11,12].

The sparse regression models can be fitted through penalized regression, also known as shrinkage methods, such as the Least·Absolute·Shrinkage and Selection·Operator (Lasso) of [13], the Smoothly·Clipped·Absolute·Deviation (SCAD) of [14], the Elastic·net (Enet) of [15], the Adaptive·Lasso of [16], the Adaptive·Enet of [17], the Minimax·Concave·Penalty (MCP) of [18], and the regression with an Elastic-SCAD (E-SCAD) of [19]. In general, these penalties are collectively referred to as folded concave penalties. However, it is interesting that shrinkage methods can attain both accurate forecasts and consistent feature selection.

The use of these methodologies along with sparse modeling has become well known because they can successfully tackle huge sets of macroeconomic data and are a noticeable alternative to factor models, as shown by [20–40]

By employing a reduction in size, the Stochastic·Dynamic·Factor (SDF, which is equivalent to large factor models) model can exhibit significant effectiveness [41,42], even when dealing with basic linear attributes from the conventional factor collection. According to [43], it may be necessary to use multiple characteristics-based parameters in order to accurately approximate the SDF. [44] provided formalization and evidence of the long-standing conjecture that, where there are a large number of characteristics-based factors, an unconditional SDF constructed from these factors will converge to the actual, conditional SDF. One can utilize Large·Factor·Models (LFMs) to construct the genuine, conditional stochastic discount factor (SDF). Although LFMs possess a high level of approximation capacity, they encounter a significant obstacle: These phenomena display significant statistical complexity and necessitate the estimation of a vast number of parameters (such as factor weights in the SDF) which greatly surpasses the number of observations. One could predict that a simplified version of the LFMs would perform better when tested with new data because it effectively reduces the problem of overfitting with the available data. [44] disproved this intuition. The studies conducted by [41] demonstrated the importance of complexity in factor pricing models. Specifically, LFMs with higher dimensions and a large number of parameters demonstrate superior performance when tested on data not used during training. These models exploit the numerous nonlinearities that are concealed in the connection between attributes and stock returns.

Similarly, ref. [45] employed a four-layer neural network consisting of 64 neurons in each layer, while ref. [46] utilized a four-layer neural network with four neurons in each layer. These narrow network topologies had a high number of parameters and functioned in regimes that were almost overfitting, as evidenced by the significant changes in their performances between training and testing data, as described in the aforementioned articles.

These characteristics render them highly challenging to analyze systematically. The loss landscape which they have is already extremely non-convex, containing many local minima and having questionable performance when applied to new data [42].

The big data environment and machine learning tools have currently garnered a great deal of attention in economic analysis [36]. When it comes to macroeconomic forecasting, ref. [29] recommended penalized regression methods; refs. [4,5,47] suggested factor-based models; and similarly, autometrics was suggested by [48]. Recently, big data was categorized by [24] into three classes—Fat·Big·Data, Huge·Big·Data, and Tall·Big·Data—which can be further illustrated as:

- Fat·Big·Data: the length of covariates (large P) exceeds the number of observations (large N);
- Tall·Big·Data: the length of covariates (large P) is considerably lower than the number of observations (sufficient large N);
- Huge·Big·Data: the length of covariates (large P) is lower than the number of observations (large N).

P and N indicate the number of covariates and the number of observations, respectively. Visually, the three types of big data are depicted in Figure 1.

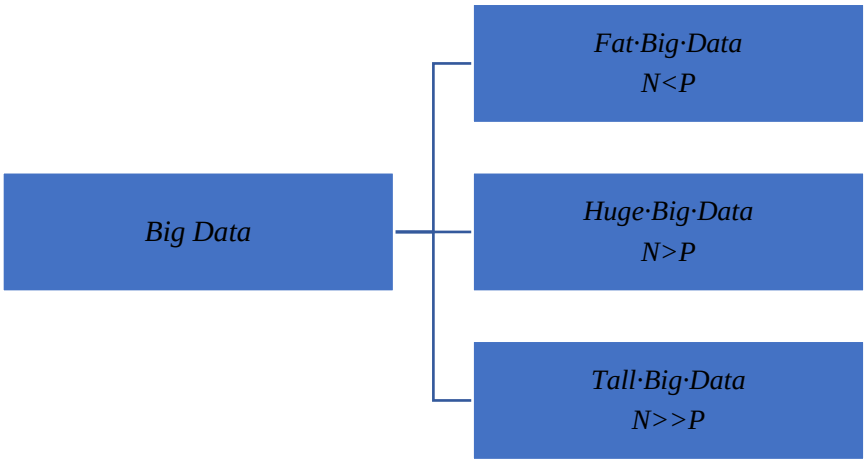


Figure 1. Categorization of big data.

Earlier research works have focused on independent component analysis, PCA, and sparse PCA for the formulation of factor-based models. However, very few past studies have used the classical method (autometrics) for time series forecasting [22] and [48,49]. Apart from this, we have not found even a single paper to date in which the forecasting performance of a factor model based on PLS analysis has been explored theoretically. Moreover, various past studies have used penalization techniques such as ridge regression, elastic net, Lasso, adaptive Lasso, and non-negative garrote, but none of the published works have yet utilized the modified versions of penalization techniques for the forecasting of macroeconomic variables.

This work employs several novel methods in big data analysis to enhance the existing empirical and theoretical research on macroeconomic forecasting by addressing the following shortcomings of a recent study which specifically concentrated on Fat·Big·Data. By utilizing dimension reduction techniques, we develop factor-based models to emphasize the impact of these models on macroeconomic forecasting. To achieve this objective, factor-based models are developed by employing PCA and PLS. In addition, we evaluate both the conventional approach and updated forms of penalization approaches, namely, MCP and E-SCAD. We provide a thorough examination of the predictive capacities of factor models, classical methods, and penalized regression techniques. To summarize the entire discussion, our primary contribution is a comparison of the forecasting performance of penalized regression tools and autometrics with factor models that have recently been

established. The comparison is constructed through the use of exhaustive simulation exercises, such as multicollinearity, autocorrelation, and heteroscedasticity, as well as empirical application to the macroeconomic dataset. The purpose of this research is to develop a more advanced tool that can be used to provide assistance to practitioners and policymakers who are working with fat big data. The improved tool is not restricted to inflation, but can be applied to any macroeconomic time series.

The remaining sections are organized as follows. Section 2 provides a thorough discussion of factor, classical and penalized methods. Simulation exercise on the comparative performance of various forecasting methods is discussed in Section 3. Empirical results and visualization are presented in Section 4. Concluding remarks are given in Section 5.

2. Methods

To effectively tackle the challenges presented by fat big data, we use a comprehensive set of advanced statistical methodologies, as well as penalization and machine learning techniques. Figure 2 depicts several approaches in great detail, including factor models based on PLS and PCA, as well as traditional econometric methods such as autometrics. In addition, we use penalization methods like Lasso, Elastic Net, and SCAD to improve predictive accuracy and model selection. Our methodology is intended to solve major concerns such as multicollinearity, heteroscedasticity, and autocorrelation, resulting in a robust and dependable forecasting performance.

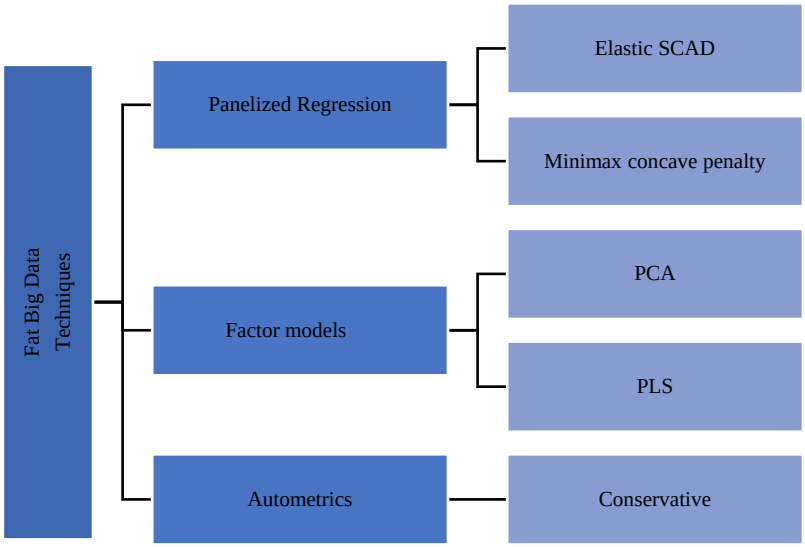


Figure 2. Schematic representation of fat big data methods.

2.1. Factor Models

One of the most widely applied methods in macroeconomic forecasting, under a large set of features, is principal component analysis, which is based on factor models suggested by [4,5]. The basic notion behind factor models is to distill the unseen, hidden factors from a huge set of features and then to utilize a relatively small number of factors as covariates for predictive modeling. Suppose Z_{it} is a potential candidate covariate generated from the following equation:

$$Z_{jk} = \pi_j' F_k^s + \epsilon_{jk} \quad (1)$$

For $j = 1, 2, \ldots, M$, and $k = 1, 2, \ldots, N$, $F_k^s = (f_{1k}, f_{2k}, \ldots, f_{lk})'$ is a vector of size 's' common factors, $\pi_j^{'}$ is a vector of size 's' factor loadings, and ϵ_{jk} denotes the idiosyncratic random term.

The PCR: The formulation of factor-based model requires the following two steps. In the first step, the F_k^s latent factors are extracted as principal components using all included

covariates Z_{jk} by minimizing the term $\sum_{j=1}^{M}\sum_{k=1}^{N}\left(Z_{jk}-\pi_j' F_k^s\right)^2$. In the second step, the h-period of the sample forecast is constructed by running the PCR as follows:

$$R_{k+h}= \gamma_j' F_k^s + e_{k+h} \tag{2}$$

where γ_j, the dimension of estimated coefficients, is 's', which is basically estimated from R_k and F_k. Detailed discussions regarding the factor approach are given by [4,30,50,51].

This method is most commonly employed in the literature on factor model, as PCs are easily generated using singular value decompositions [4,52,53].

It is more likely that the factor approach will provide a poor forecast if the included common factors are dominated by omitted common factors [54]. Likewise, ref. [34] argued that PCA utilizes the factor structure for Z and does not account for the response variable. It illustrates, no matter what, that the response is variable in the forecast. Due to ignoring the response variable during factors extraction, the resulting model's forecast is inaccurate.

The PLS method: This study takes into account PLS regression, a widely used alternative to PCA which was first introduced by [55]. The approach is suitable in a mountain-of-data environment (fat big data) and is deemed an alternative to factor models constructed using PCA. In contrast to PCA, PLS yields independent components by utilizing the existing association amid covariates and the corresponding response variable, although it also retains most of the variance of the covariates. PLS has proved to be successful in situations where the number of predictors (P) is sufficiently larger than the number of data points (N) and extreme multicollinearity exists among the covariates [56]. Generally, the PLS approach seeks the directions of maximal variance that assist in delineating the covariates, as well as the response variable. The mathematical form of the PLS can be expressed as

$$R_t= y_t\alpha_P+\varepsilon_t \tag{3}$$

where $y_t = [y_{1,t}, y_{2,t}, \ldots, y_{k,t}]'$ is a vector of covariates of size $k \times 1$ observed at time $t = 1, \ldots, T$; α_P is a vector of coefficients with a dimension $k \times 1$; and ε_t is a random error. To achieve a k-period ahead of the sample forecast, we may utilize the equation given below:

$$\hat{R}_{t+k}= \hat{\alpha}_k' y_t \tag{4}$$

2.2. Panelized Regression and Classical Approach

In addition to the above factor models, we also consider methods of penalized regression, including MCP and E-SCAD, as well as the classical method (autometrics), as both approaches are good alternatives to factor models. Here, we give concise outlines of a number of these approaches, as well as the corresponding citations to thorough debates regarding them.

Panelized·Regression·Methods

The parameters of the included Panelized·Regression·Methods are estimated according to the following objective function:

$$\sum_{t=1}^{T}\left\{\left(R_{k+h}-\sum_{i=1}^{n}\alpha_i y_{it}\right)^2+\pi g(\alpha)\right\} \tag{5}$$

where π refers to the hyperparameter of regularization. The specification of a penalty term $g(\alpha)$ differs for the aforementioned penalized techniques; by definition, α is equal to $(\alpha_1, \alpha_2, \ldots, \alpha_n)'$. For the selection of hyperparameter π, we adopt a cross-validation approach in our study, following [36].

MCP: The MCP was initially developed by [18]. It corresponds to the penalized family of regression with a penalty term $g_\pi(\alpha) = \frac{(\aleph\pi-\vartheta)+}{\aleph}$. According to Zhang, the probability that the MCP penalty may choose the right model tends to be 1. Moreover, in terms of Lq-loss, the MCP estimator enjoys oracle properties provided that $\aleph$ and π ensure certain

conditions in a high-dimensional setting [57]. More recently, the MCP has shown very interesting findings in terms of variable selection, estimation, and forecasting [58].

E-SCAD: SCAD was modified by adding the L_2 penalty. The new method is called elastic SCAD (E-SCAD). In addition to an oracle property, E-SCAD achieves an extra property in which the penalty function drives the inclusion or exclusion of a strongly correlated set of predictors from the model. To accomplish this, the process does not require any prior information [19]. Mathematically, the penalty function of E-SCAD is given as follows:

$$g_\pi(\alpha) = \sum_{c=1}^{C} g_\pi(\alpha_c) + \lambda_2 \sum_{c=1}^{n} \alpha_d^2 \tag{6}$$

2.3. Classical Approach

Autometrics is a popular statistical approach which is applicable in the case of huge big data as well as fat big data [24]. In general, the algorithm of autometrics basically consists of five steps. In the initial step, the model is designed in a linear form in which all the covariates are included, called a Generalized·Unrestricted·Model (GUM). The second step provides us with the estimates of unknown parameters and tests them for statistical significance. The third step involves the pre-search process and is followed by a tree path search in step four. In step five, the model is selected for forecasting.

We obtain the forecasting model by implementing autometrics into the GUM:

$$R_t = \beta_0 + \sum_{u=1}^{n} \sum_{v=1}^{m} \alpha_{u,v} y_{u,t-v} + \varepsilon_t \tag{7}$$

For model selection, the liberal strategy, also known as the super-conservative strategy, is considered in this study. This strategy is primarily based on a level of significance of one percent. Put differently, the significance of the estimated coefficients is based on a significance level of one percent.

3. Monte Carlo Evidence on Forecasting Performance

This section performs some simulation exercises intending to explore the predictive power of factor models against classical and penalization methods. In doing so, we consider three main scenarios: multicollinearity, heteroscedasticity, and autocorrelation. Considering the cases of multicollinearity, three types of correlation structure among the set of features are assumed—low (0.25), moderate (0.50), and high (0.90)—under the normally distributed errors. To generate the artificial data for our simulation experiments, we follow the data generation process of [24,59].

3.1. Data·Generating·Process (DGP)

The following equation is used to generate data:

$$R_i = y_i^T \alpha + \varepsilon_i \tag{8}$$

The set of covariates y_i is generated from a multivariate normal distribution with a mean of zero, and the pairwise covariance between m and n is $cov(x_m, x_n) = \sum^{|m-n|}$ [59]. We split the two candidate sets of variables into 50 and 70, then further divide them into relevant (p) and irrelevant (q) variables, as depicted in Figure 3.

The second scenario explores the forecast performance in the presence of autocorrelation. More specifically, this refers to how factor models compete with the rival methods provided the error term of a model is autocorrelated. The correlation between current and lagged realizations is symbolized by ρ, which is generated as

$$\varepsilon_t = \rho \varepsilon_{t-1} + \mu_t \tag{9}$$

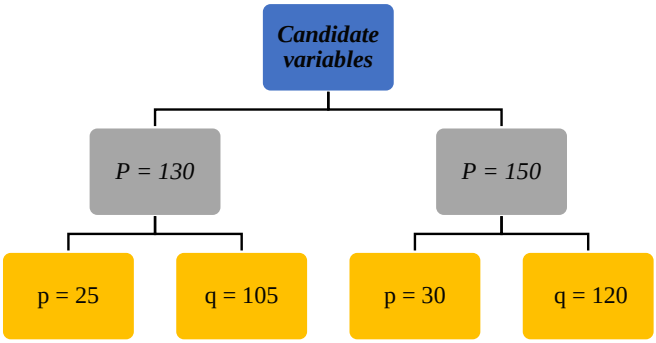


Figure 3. Classification of candidate variables into p and q variables.

Our simulation exercises assume three levels of autocorrelation—low, moderate, and high—as, for example, $\rho \in \{0.25, 0.5, 0.9\}$.

Similarly, the third scenario focuses on heteroscedasticity, which demonstrates the variance of the error term across observations by δ_k.

$$\mathrm{E}\left(\mu_t^2\right) = \delta_k \tag{10}$$

Thus, we break the variance δ_k into two segments, i.e., δ_1 and δ_2. Suppose there are 'n' data points, we adjust the variance of (n/2) data points as δ_1 and the remaining data points variance to be δ_2. Our simulation exercises conjecture the low, moderate, and high levels of heteroscedasticity and adjust the values of $\pi_i = (\sigma_1/\sigma_2)$, where i = 1, 2, 3 and $\pi_i \in \{0.1/0.3, 0.2/0.6, 0.3/0.9\}$. For all penalization techniques and factor models in our study, we select the optimal hyper parameter(s) by means of tenfold cross-validation

We divide the dataset so that 80 percent of the data are used for model training and the remaining data are used for model evaluation in order to compare the prediction capabilities of all procedures. We repeat the process H = 1000 times. The mean of the Root·Mean·Square·Error (RMSE) and the Mean·Absolute·Error (MAE) are calculated over 'H' to evaluate the predictive power. Through these two criteria, we can achieve prediction accuracy with all included methods. The smaller values of MAE and RMSE indicate comparatively better forecasts. To obtain the simulation and empirical results, we rely on various packages, like gets, pls, caret, ncvreg, Metrics, and forecast, under the R programming language.

3.2. Simulation Results

The forecast comparison output obtained from Monte Carlo exercises is reported in Tables 1–3. The entries in bold show the best performances of the underlying model. It can be observed that the performances of all procedures improve with the increasing data points.

Scenario 1. Considering the cases of low and moderate multicollinearity, the predictive ability of autometrics is more effective than that of competing methods. But, in case of a small sample, the RMSE and MAE associated with autometrics are slightly better than the PLS-based factor approach. This clearly indicates that the PLS-based factor approach is strongly competitive. Similarly, despite achieving a considerable improvement in RMSE and MAE by E-SCAD when the sample size is increased, the results are not satisfactory in contrast to autometrics. Moreover, by increasing the number of active and inactive variables, autometrics remains dominant, with the lowest RMSE and MAE. In the presence of extreme multicollinearity, the factor approach based on PLS outperforms its rival counterparts in terms of the lowest forecast error. According to both error criteria, autometrics stands as a good competing method.

Scenario 2. Based on RMSE and MAE, the forecasting capabilities of autometrics are superior to those of all competing counterparts in the presence of heteroscedasticity. In contrast, the MCP and E-SCAD perform poorly using a small size, but as we expand the

data window (large sample), their forecasting performance dramatically improves. This indicates that penalized regression models require a large number of data points in order to provide accurate forecasts.

Scenario 3. Despite adding more irrelevant variables, autometrics demonstrates remarkable forecasting performance for low and moderate autocorrelation. E-SCAD remains a good competing method, particularly when more observations are used. Considering the extreme autocorrelation, E-SCAD provided the lowest RMSE and MAE compared to the other competing counterparts, but autometrics was still a good contester.

Table 1. Variable selection under multicollinearity from Monte Carlo simulation.

Models	**$\sum$ = 0.25, *P* = 130**		**$\sum$ = 0.25, *P* = 150**	
n = 50/100/125	**RMSE**	**MAE**	**RMSE**	**MAE**
MCP	6.86/5.41/2.243	5.602/4.375/1.811	6.043/3.319/1.208	4.970/2.681/0.975
E-SCAD	5.80/2.00/1.355	4.741/1.620/1.095	4.899/1.364/1.257	4.008/1.098/1.016
Autometrics	**4.192/1.312/1.189**	**3.419/1.058/0.957**	**3.267/1.222/1.145**	**2.673/0.986/0.924**
PLS_FM	4.530/3.213/2.727	3.678/2.589/2.197	5.260/3.786/3.295	4.309/3.062/2.623
PCA_FM	6.475/5.781/5.695	5.725/4.685/4.589	6.512/6.398/6.342	5.318/5.166/5.104
n = 50/100/125	**$\sum$ = 0.50, *P* = 130**		**$\sum$ = 0.50, *P* = 150**	
MCP	7.918/4.414/3.007	6.505/3.564/2.429	6.512/3.192/1.748	5.353/2.579/1.406
E-SCAD	5.380/2.093/1.581	4.414/1.688/1.276	4.118/1.548/1.326	3.375/1.247/1.070
Autometrics	**4.394/1.469/1.221**	**3.282/1.186/0.983**	**3.178/1.325/1.159**	**2.599/1.069/0.934**
PLS_FM	4.414/2.533/2.151	3.60/2.043/1.732	5.285/3.037/2.519	4.330/2.458/2.029
PCA_FM	6.724/6.310/6.186	5.544/5.107/4.977	7.809/7.255/7.076	6.402/5.854/5.698
n = 50/100/125	**$\sum$ = 0.90, *P* = 130**		**$\sum$ = 0.90, *P* = 150**	
MCP	5.031/3.784/3.638	4.101/3.057/2.932	4.123/3.253/3.146	3.372/2.636/2.541
E-SCAD	2.699/2.344/2.307	2.215/1.895/1.856	2.222/2.024/2.016	1.817/1.630/1.629
Autometrics	2.709/1.982/1.757	2.219/1.605/1.418	2.437/1.788/1.620	2.001/1.443/1.307
PLS_FM	**1.797/1.347/1.274**	**1.472/1.086/1.027**	**2.080/1.426/1.326**	**1.706/1.143/1.069**
PCA_FM	3.125/2.306/2.149	2.571/1.865/1.742	4.037/2.881/2.685	3.293/2.326/2.162

Noted: Bold values show a better forecast.

Table 2. Variable selection under heteroscedasticity from Monte Carlo simulation.

Models	**π_1 = 0.1/0.3, *P* = 130**		**π_1 = 0.1/0.3, *P* = 150**	
n = 50/100/125	**RMSE**	**MAE**	**RMSE**	**MAE**
MCP	6.317/2.072/0.472	5.183/1.679/0.381	7.656/3.935/1.649	6.244/3.178/1.331
E-SCAD	3.824/0.849/0.668	3.131/0.686/0.539	5.143/1.412/0.948	4.194/1.145/0.765
Autometrics	**0.403/0.327/0.317**	**0.330/0.264/0.255**	**0.582/0.332/0.326**	**0.477/0.268/0.263**
PLS_FM	4.236/1.985/1.328	3.455/1.603/1.070	5.146/2.668/1.898	4.216/2.158/1.530
PCA_FM	6.658/6.222/6.134	5.477/5.037/4.936	7.863/7.195/7.043	6.300/5.805/5.668
n = 50/100/125	**π_2 = 0.2/0.6, *P* = 130**		**π_2 = 0.2/0.6, *P* = 150**	
MCP	6.419/2.349/0.798	5.269/1.899/0.642	7.711/4.002/1.962	6.296/3.238/1.583
E-SCAD	3.891/1.038/0.871	3.185/0.837/0.703	5.186/1.567/1.121	4.222/1.270/0.906
Autometrics	**0.974/0.653/0.644**	**0.798/0.528/0.519**	**1.765/0.668/0.653**	**1.443/0.538/0.527**
PLS_FM	4.277/2.106/1.555	3.487/1.695/1.253	5.178/2.743/2.038	4.485/2.220/1.645
PCA_FM	6.680/6.244/6.155	5.495/5.050/4.952	7.735/7.216/7.055	6.350/5.822/5.679
n = 50/100/125	**π_3 = 0.3/0.9, *P* = 130**		**π_3 = 0.3/0.9, *P* = 150**	
MCP	6.463/2.661/1.152	5.300/2.147/0.926	7.743/4.131/2.292	6.363/3.339/1.851
E-SCAD	3.983/1.293/1.131	3.263/1.043/0.912	5.257/1.796/1.359	4.287/1.455/1.097
Autometrics	**1.939/0.977/0.958**	**1.588/0.785/0.772**	**2.730/1.010/0.975**	**2.225/0.818/0.786**
PLS_FM	4.345/2.281/1.838	3.542/1.839/1.480	5.234/2.867/2.241	4.292/2.320/1.807
PCA_FM	6.719/6.284/6.203	5.540/5.084/4.991	7.780/7.241/7.089	6.386/5.843/5.705

Noted: Bold values show a better forecast.

Table 3. Variable selection under autocorrelation from Monte Carlo simulation.

Models	ρ = 0.25, P = 130		ρ = 0.25, P = 150	
n = 50/100/124	**RMSE**	**MAE**	**RMSE**	**MAE**
MCP	6.566/3.266/1.780	5.364/2.641/1.440	7.935/4.379/3.076	6.488/3.541/2.475
E-SCAD	4.254/1.614/1.364	3.475/1.306/1.102	5.335/2.154/1.609	4.362/1.738/1.297
Autometrics	**3.253/1.407/1.214**	**2.659/1.137/0.982**	**4/1.548/1.278**	**3.248/1.252/1.030**
PLS_FM	4.520/2.617/2.204	3.695/2.117/1.777	5.330/3.096/2.538	4.348/2.499/2.048
PCA_FM	6.713/6.282/6.195	5.490/5.073/4.987	7.691/7.239/7.024	6.291/5.869/5.697
n = 50/100/124	ρ = 0.50, P = 130		ρ = 0.50, P = 150	
MCP	6.642/3.376/2.111	5.441/2.722/1.702	7.996/4.524/3.295	6.562/3.654/2.663
E-SCAD	4.326/1.756/1.507	3.541/1.422/1.220	5.359/2.310/1.772	4.364/1.867/1.431
Autometrics	**3.462/1.622/1.388**	**2.840/1.316/1.123**	**4.470/1.789/1.489**	**3.637/1.446/1.201**
PLS_FM	4.585/2.689/2.329	3.764/2.174/1.877	5.330/3.207/2.683	4.362/2.598/2.164
PCA_FM	6.847/6.393/6.228	5.611/5.142/5.019	7.457/7.214/7.185	6.108/5.853/5.796
n = 50/100/124	ρ = 0.90, P = 130		ρ = 0.90, P = 150	
MCP	7.069/4.646/4.002	5.771/3.782/3.257	8.268/5.544/4.678	6.780/4.84/3.781
E-SCAD	**4.963/3.279/2.923**	**4.065/2.705/2.425**	**5.957/3.653/3.193**	**4.901/3.001/2.623**
Autometrics	5.257/3.687/3.329	0.268/3.031/2.736	6.169/4.013/3.573	5.013/3.270/2.916
PLS_FM	5.128/3.822/3.454	4.209/3.129/2.822	7.735/7.216/7.035	6.350/5.822/5.679
PCA_FM	6.939/6.692/6.601	5.664/5.426/5.322	7.964/7.523/7.459	6.530/6.079/6.015

Noted: Bold values indicate a better forecast.

4. Testing on Empirical Dataset

Complementing the simulation exercises, we analyze the macroeconomic time series dataset for Pakistan.

The dataset consists of 79 aggregated and dis-aggregated variables collected at a monthly frequency during the period from 2013 to 2020. The dataset covers the fiscal sector, real sector, financial and monetary sector, and external sector of the economy of Pakistan. The data are taken from the state bank of Pakistan. The forecasting model is constructed for inflation (INF). For this model, a long list of predictor variables is selected. All the variables are transformed in order to make them stationary prior to empirical analysis. Generally, logarithmic transformation is performed for all non-negative time series that are not already in rates [5]. A complete list of variables is given in Appendix A. Table A1 (given in Appendix A) contains information on the variables utilized in the analysis.

Out-of-Sample Inflation Forecasting

The time series is divided into two parts (shown by dashed line) in order to facilitate out-of-sample forecast accuracy, as shown in Figure 4. For model estimation, we utilize the data from January 2013 to February 2019 and March 2019 to December 2020 to assess the models' post-sample prediction accuracy multiple steps ahead.

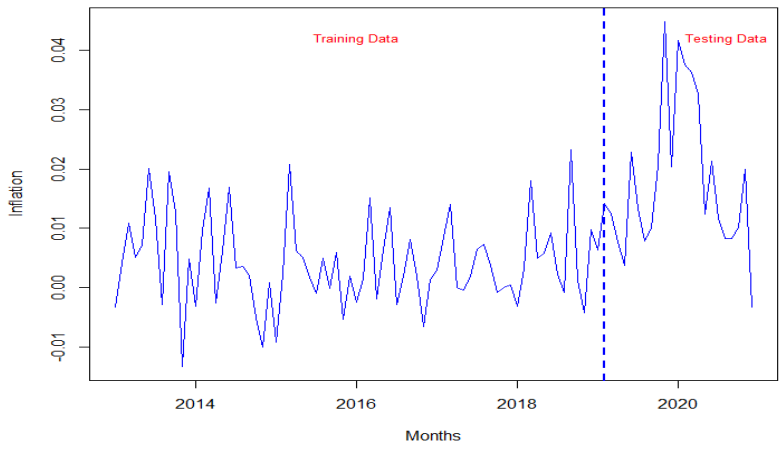
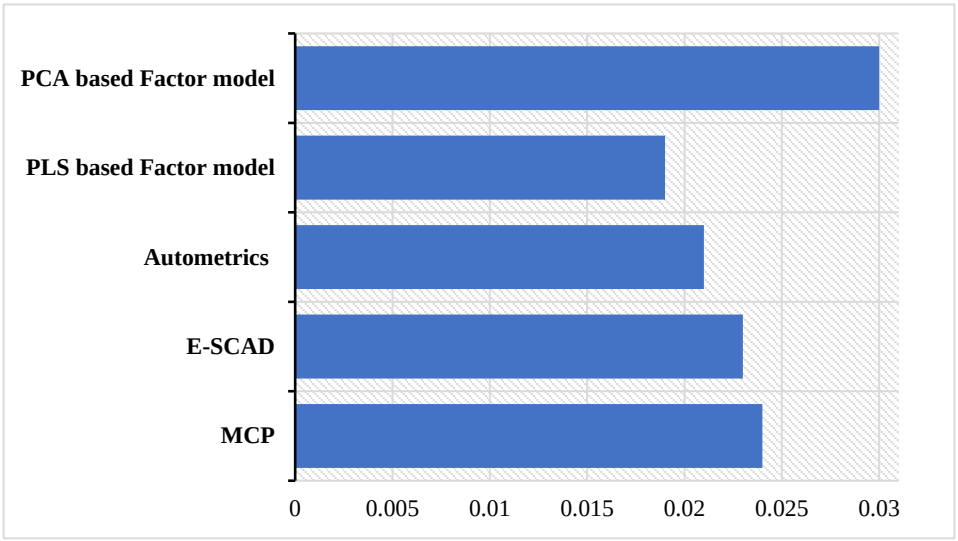
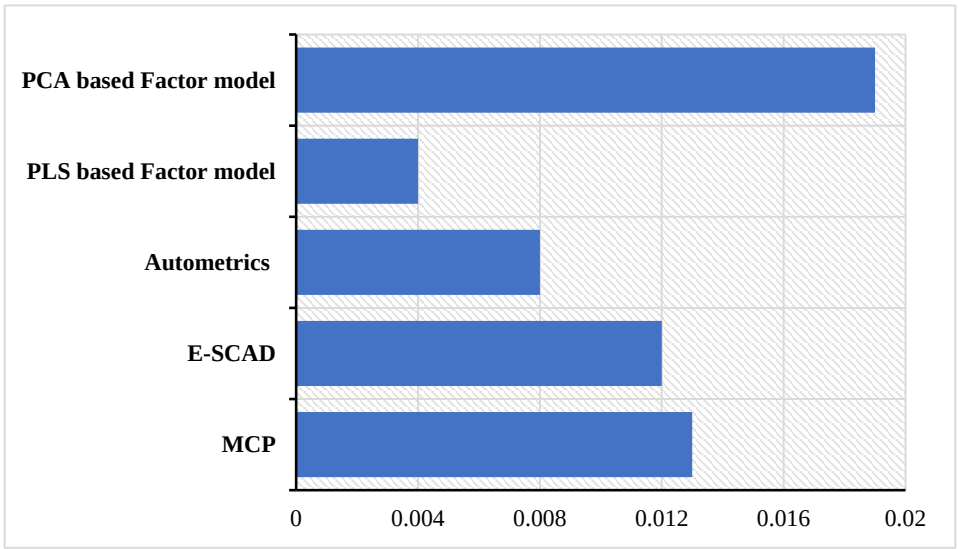


Figure 4. Monthly inflation series against time.

Figure 5a,b present the forecasting experiment across different forecasting methods for one of the core macroeconomic variables of interest (inflation). The forecasting accuracy is given as the RMSE and MAE, represented in our case by a bar chart showing the results of different methods. The smaller the length of a bar, the better the forecast attained by the model, comparatively. By observing the length of a bar given in Figure 5a,b, we can infer that the PLS-based factor model was more superior to its rival counterparts in the post-sample forecast. In contrast, the autometrics produced a good forecast, but it was not as satisfactory as that provided by the PLS-based factor model.



(**a**)



(**b**)

Figure 5. Out-of-sample forecast comparison. PLS based factor model outperforms the competing methods.

5. Discussion, Implications, and Limitations

In this section, we provide a discussion and explore the implications and limitations of the study. We evaluate the prediction capacities of several statistical models and machine learning technologies in time series forecasting scenarios using both theoretical examination and empirical research.

5.1. Discussion

This article explores the predictive power of widely used statistical models against classical and sophisticated machine learning tools theoretically as well as empirically. To be more specific, our core aim was to discover how well the most popular models in the context of time series forecasting, that is, factor models, performed against classical and shrinkage methods. Different sample sizes and predictor variables were used to evaluate each technique under the conditions of multicollinearity, heteroscedasticity, and autocorrelation. Across the simulation exercises, it was found that all methods were consistent. In the presence of low and moderate multicollinearity, based on RMSE and MAE values, autometrics outperformed the other competitive counterparts. Considering the extreme case of multicollinearity, the PLS-based factor approach beat the rival counterparts, as it had the lowest forecast error. Considering different levels of heteroscedasticity, the lowest RMSE and MAE values were attained by autometrics, which indicates its dominance over all other methods in post-sample forecasting. Across low and moderate levels of autocorrelation, autometrics produced a better forecast, but in contrast, the E-SCAD provided the lowest RMSE and MAE values for extreme autocorrelation.

5.2. Implication

Complementing the simulation exercise, we carried out an empirical application on a well-known Pakistan macroeconomic dataset. The dataset entailed 79 time series observed at a monthly frequency from January 2013 to December 2020, and was collected from the state bank of Pakistan. For model estimation, we utilized data from January 2013 to February 2019 and March 2019 to December 2020 for evaluating the models' post-sample forecasting accuracy multiple steps ahead. The statistical accuracy measures, namely, RMSE and MAE, were used in order to compare the post-sample predictive ability of the factor models against autometrics and ML techniques. Based on both statistical measures, the factor approach derived from PLS produced a better forecast than the competing counterparts. These results are consistent with the findings of [59].

5.3. Limitations and Future Avenue

There are several limitations of this study. First, it concentrated merely on linear models and was confined to monthly data. Moreover, the simulation exercise was confined to normally distributed errors, but in general, this would not be the case in a real-world phenomenon. Hence, future work can be carried out to fill the preceding research's gaps.

Author Contributions: Conceptualization, O.A.; Methodology, F.K.; Software, F.K.; Formal analysis, F.K.; Data curation, F.K.; Visualization, F.K.; Supervision, O.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Details of variables.

Sr. no	Name of the Variables
	Real Sector (Output)
1	Production of Sugar (SA)
2	Production of Vegetable (SA)
3	Production of Cigarettes (SA)

Table A1. *Cont.*

Sr. no	Name of the Variables
4	Production of Cotton yarn (SA)
5	Production of Cotton Cloth (SA)
6	Production of Paper (SA)
7	Production of Paper Board (SA)
8	Production of Soda Ash (SA)
9	Production of Caustic Soda (SA)
10	Production of Sulfuric Acid (SA)
11	Production of Chlorine Gas (SA)
12	Production of Urea (SA)
13	Production of Super Phosphate (SA)
14	Production of Ammonium Nitrate (SA)
15	Production of Nitro Phosphate (SA)
16	Production of Cycle Tyres and Tubes (SA)
17	Production of Motor Tyres and Tubes (SA)
18	Production of Cement (SA)
19	Production of Tractors (SA)
20	Production of Bicycle (SA)
21	Production of Silica Sand (SA)
22	Production of Gypsum (SA)
23	Production of Limestone (SA)
24	Production of Rock Salt (SA)
25	Production of Coal (SA)
26	Production of Chromate (SA)
27	Production of Crude Oil (SA)
28	Production of Natural Gas (SA)
29	Production of Electricity (SA)
	Monetary Sector (Money, Reserves and Banking System) **Money**
30	Currency in circulation
31	Bank Deposit with State Bank of Pakistan
32	Other Deposit with State Bank of Pakistan
33	Currency in Tills of Scheduled Banks
34	Demand Deposits
35	Time Deposits
36	Resident Foreign Currency Deposits
37	Government Sector Borrowing (net)
38	Budgetary Support
39	Commodity Operations
40	Credit to Private Sector
41	Credit to Public Sector Enterprises
42	Net Foreign (Domestic) Assets of State Bank of Pakistan
43	Net Foreign Assets of the Scheduled Banks in Pakistan

Table A1. *Cont.*

Sr. no	Name of the Variables
	Prices
44	Consumer Price Index
45	Consumer Price Index (Food)
46	Wholesale Price Index
47	Sensitive Price Index
	Exchange Rates
48	Nominal Effective Exchange Rate
49	Real Effective Exchange Rate
50	Saudi Arabian Riyal (Monthly Average)
51	UAE Dirham (Monthly Average)
52	US Dollar (Monthly Average)
53	Canadian Dollar (Monthly Average)
54	UK Pound Sterling (Monthly Average)
55	Euro (Monthly Average)
56	Japanese Yen (Monthly Average)
	Interest Rates
57	Lending Weighted Average Rates
58	Deposits Weighted Average Rates
59	Call Money Rate
60	Overnight Weighted Average Repo Rate (all data)
61	Karachi Interbank Offered Rate 1 Week
62	Karachi Interbank Offered Rate 2 Weeks
63	Karachi Interbank Offered Rate 1 Month
64	Karachi Interbank Offered Rate 3 Months
65	Karachi Interbank Offered Rate 6 Months
66	Karachi Interbank Offered Rate 9 Months
67	Karachi Interbank Offered Rate 12 Months
	External Sector
68	Exports
69	Imports
70	Workers Remittances
71	Gold Reserves
72	Foreign Exchange Reserves with State Bank of Pakistan
73	Foreign Exchange Reserves with Scheduled Banks in Pakistan
74	Old Foreign Currency Accounts
75	New Foreign Currency Accounts (FE-25)
	Fiscal Sector
76	Federal Government Direct Tax Collection
77	Federal Government Indirect Tax (Sales Tax)
78	Federal Government Indirect Tax (Excise Tax)
79	Federal Government Indirect Tax (Customs)

References

1. Filzmoser, P.; Nordhausen, K. Robust linear regression for high-dimensional data: An overview. *Wiley Interdiscip. Rev. Comput. Stat.* **2021**, *13*, e1524. [CrossRef]
2. Gujarati, D.N.; Porter, D.C.; Gunasekar, S. *Basic Econometrics*; Tata McGraw-Hill Education: New York, NY, USA, 2012.
3. Kim, H.H.; Swanson, N.R. *Mining Big Data Using Parsimonious Factor and Shrinkage Methods*; Working paper; Rutgers University: New Brunswick, NJ, USA, 2013.
4. Stock, J.H.; Watson, M.W. Macroeconomic forecasting using diffusion indexes. *J. Bus. Econ. Stat.* **2002**, *20*, 147–162. [CrossRef]
5. Stock, J.H.; Watson, M.W. Generalized shrinkage methods for forecasting using many predictors. *J. Bus. Econ. Stat.* **2012**, *30*, 481–493. [CrossRef]
6. Hansen, C.; Liao, Y. The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications. *Econom. Theory* **2019**, *35*, 465–509. [CrossRef]
7. Bai, J.; Liao, Y. Efficient estimation of approximate factor models via penalized maximum likelihood. *J. Econom.* **2016**, *191*, 1–18. [CrossRef]
8. Fan, J.; Ke, Y.; Liao, Y. Robust factor models with explanatory proxies. *arXiv* **2016**, arXiv:1603.07041. [CrossRef]
9. Fan, J.; Liao, Y.; Wang, W. Projected principal component analysis in factor models. *Ann. Stat.* **2016**, *44*, 219. [CrossRef]
10. Fan, J.; Xue, L.; Yao, J. Sufficient forecasting using factor models. *J. Econom.* **2017**, *201*, 292–306. [CrossRef]
11. Bernanke, B.S.; Boivin, J.; Eliasz, P. Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *Q. J. Econ.* **2005**, *120*, 387–422.
12. Syed, A.A.S.; Lee, K.H. Macroeconomic forecasting for Pakistan in a data-rich environment. *Appl. Econ.* **2021**, *53*, 1077–1091. [CrossRef]
13. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
14. Fan, J.; Li, R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
15. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320. [CrossRef]
16. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
17. Zou, H.; Zhang, H.H. On the adaptive elastic-net with a diverging number of parameters. *Ann. Stat.* **2009**, *37*, 1733. [CrossRef] [PubMed]
18. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef] [PubMed]
19. Zeng, L.; Xie, J. Group variable selection via SCAD-L 2. *Statistics* **2014**, *48*, 49–66. [CrossRef]
20. Bai, J.; Ng, S. Forecasting economic time series using targeted predictors. *J. Econom.* **2008**, *146*, 304–317. [CrossRef]
21. De Mol, C.; Giannone, D.; Reichlin, L. Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *J. Econom.* **2008**, *146*, 318–328. [CrossRef]
22. Castle, J.L.; Clements, M.P.; Hendry, D.F. Forecasting by factors, by variables, by both or neither? *J. Econom.* **2013**, *177*, 305–319. [CrossRef]
23. Luciani, M. Forecasting with approximate dynamic factor models: The role of non-pervasive shocks. *Int. J. Forecast.* **2014**, *30*, 20–29. [CrossRef]
24. Doornik, J.A.; Hendry, D.F. Statistical model selection with big data. *Cogent Econ. Financ.* **2015**, *3*, 1045216. [CrossRef]
25. Kristensen, J.T. Diffusion indexes with sparse loadings. *J. Bus. Econ. Stat.* **2017**, *35*, 434–451. [CrossRef]
26. Li, J.; Chen, W. Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models. *Int. J. Forecast.* **2014**, *30*, 996–1015. [CrossRef]
27. Marsilli, C. Variable Selection in Predictive MIDAS Models. Banque de France Working Paper No. 520. 2014. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2531339 (accessed on 17 June 2024).
28. Nicholson, W.; Matteson, D.; Bien, J. BigVAR: Tools for modeling sparse high-dimensional multivariate time series. *arXiv* **2017**, arXiv:1702.07094.
29. Kim, H.H.; Swanson, N.R. Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *J. Econom.* **2014**, *178*, 352–367. [CrossRef]
30. Kim, H.H.; Swanson, N.R. Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *Int. J. Forecast.* **2018**, *34*, 339–354. [CrossRef]
31. Swanson, N.R.; Xiong, W. Big data analytics in economics: What have we learned so far, and where should we go from here? *Can. J. Econ.* **2018**, *51*, 695–746. [CrossRef]
32. Swanson, N.R.; Xiong, W.; Yang, X. Predicting interest rates using shrinkage methods, real-time diffusion indexes, and model combinations. *J. Appl. Econom.* **2020**, *35*, 587–613. [CrossRef]
33. Smeekes, S.; Wijler, E. Macroeconomic forecasting using penalized regression methods. *Int. J. Forecast.* **2018**, *34*, 408–430. [CrossRef]
34. Tu, Y.; Lee, T.H. Forecasting using supervised factor models. *J. Manag. Sci. Eng.* **2019**, *4*, 12–27. [CrossRef]
35. Kim, H.; Ko, K. Improving forecast accuracy of financial vulnerability: PLS factor model approach. *Econ. Model.* **2020**, *88*, 341–355. [CrossRef]

36. Maehashi, K.; Shintani, M. Macroeconomic forecasting using factor models and machine learning: An application to Japan. *J. Jpn. Int. Econ.* **2020**, *58*, 101104. [CrossRef]
37. Abdić, A.; Resić, E.; Abdić, A. Modelling and forecasting GDP using factor model: An empirical study from Bosnia and Herzegovina. *Croat. Rev. Econ. Bus. Soc. Stat.* **2020**, *6*, 10–26. [CrossRef]
38. Kim, H.; Shi, W.; Kim, H.H. Forecasting financial stress indices in Korea: A factor model approach. *Empir. Econ.* **2020**, *59*, 2859–2898. [CrossRef]
39. Kim, H.; Shi, W. Forecasting financial vulnerability in the USA: A factor model approach. *J. Forecast.* **2021**, *40*, 439–457. [CrossRef]
40. Khan, F.; Urooj, A.; Khan, S.A.; Alsubie, A.; Almaspoor, Z.; Muhammadullah, S. Comparing the Forecast Performance of Advanced Statistical and Machine Learning Techniques Using Huge Big Data: Evidence from Monte Carlo Experiments. *Complexity* **2021**, *2021*, 6117513. [CrossRef]
41. Kelly, B.T.; Kuznetsov, B.; Malamud, S.; Xu, T.A. *Deep Learning from Implied Volatility Surfaces*; Swiss Finance Institute Research Paper; Swiss Finance Institute: Zürich, Switzerland, 2023; pp. 23–60.
42. Kelly, B.; Kuznetsov, B.; Malamud, S.; Xu, T.A. Large (and Deep) Factor Models. *arXiv* **2024**, arXiv:2402.06635. [CrossRef]
43. Kozak, S.; Nagel, S. *When Do Cross-Sectional Asset Pricing Factors Span the Stochastic Discount Factor? (No. w31275)*; National Bureau of Economic Research: Cambridge, MA, USA, 2023.
44. Didisheim, A.; Ke, S.B.; Kelly, B.T.; Malamud, S. *Complexity in Factor Pricing Models (No. w31689)*; National Bureau of Economic Research: Cambridge, MA, USA, 2023.
45. Chen, L.; Pelger, M.; Zhu, J. Deep learning in asset pricing. *Manag. Sci.* **2024**, *70*, 714–750. [CrossRef]
46. Fan, J.; Ke, Z.T.; Liao, Y.; Neuhierl, A. Structural Deep Learning in Conditional Asset Pricing. Available at SSRN 4117882. 2022. Available online: https://static1.squarespace.com/static/5d6417169b0edd0001903770/t/655524542cbf566e3801a2ed/1700078678513/guilherme+piancetino.pdf (accessed on 17 June 2024).
47. Stock, J.H.; Watson, M.W. Forecasting inflation. *J. Monet. Econ.* **1999**, *44*, 293–335. [CrossRef]
48. Castle, J.L.; Doornik, J.A.; Hendry, D.F. Modelling non-stationary 'Big Data'. *Int. J. Forecast.* **2021**, *37*, 1556–1575. [CrossRef]
49. Khan, F.; Urooj, A.; Khan, S.A.; Khosa, S.K.; Muhammadullah, S.; Almaspoor, Z. Evaluating the performance of feature selection methods using huge big data: A Monte Carlo simulation approach. *Math. Probl. Eng.* **2022**, *2022*, 6607330. [CrossRef]
50. Stock, J.H.; Watson, M.W. Forecasting using principal components from a large number of predictors. *J. Am. Stat. Assoc.* **2002**, *97*, 1167–1179. [CrossRef]
51. Bai, J.; Ng, S. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* **2006**, *74*, 1133–1150. [CrossRef]
52. Bai, J.; Ng, S. Determining the number of factors in approximate factor models. *Econometrica* **2002**, *70*, 191–221. [CrossRef]
53. Bai, J.; Ng, S. Evaluating latent and observed factors in macroeconomics and finance. *J. Econom.* **2006**, *131*, 507–537. [CrossRef]
54. Boivin, J.; Ng, S. Are more data always better for factor analysis? *J. Econom.* **2006**, *132*, 169–194. [CrossRef]
55. Wold, H. *Soft Modelling: The Basic Design and Some Extensions, Vol. 1 of Systems under Indirect Observation, Part II*; North-Holland: Amsterdam, The Netherlands, 1982.
56. Pascual Herrero, H. Least Squares Regression Principal Component Analysis. Bachelor's Thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, 2020.
57. Wang, Y.; Fan, Q.; Zhu, L. Variable selection and estimation using a continuous approximation to the L_0 penalty. *Ann. Inst. Stat. Math.* **2018**, *70*, 191–214. [CrossRef]
58. Li, N.; Yang, H. Nonnegative estimation and variable selection under minimax concave penalty for sparse high-dimensional linear regression models. *Stat. Pap.* **2021**, *62*, 661–680. [CrossRef]
59. Khan, F.; Urooj, A.; Ullah, K.; Alnssyan, B.; Almaspoor, Z. A Comparison of Autometrics and Penalization Techniques under Various Error Distributions: Evidence from Monte Carlo Simulation. *Complexity* **2021**, *2021*, 9223763. [CrossRef]