EVALUATING THE PERFORMANCE OF VARIABLE SELECTION AND FORECASTING METHODS USING BIG DATA: A MONTE CARLO SIMULATION APPROACH

Submitted By

Faridoon Khan

Student's Registration Number

PIDE2018FPHDETS04

Supervisor

Dr. Amena Urooj

Co-Supervisor

Dr. Saud Ahmed Khan

PIDE School of Economics

Pakistan Institute of Development Economics, Islamabad

2023

Author's Declaration

I Mr. Faridoon Khan hereby state that my PhD thesis titled **"Evaluating The Performance of Variable Selection and Forecasting Methods Using Big Data: A Monte Carlo Simulation Approach"** is my own work and has not been submitted previously by me for taking any degree from **Pakistan Institute of Development Economics, Islamabad'** or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my Graduation the university has the right to withdraw my PhD degree.

Mr. Faridoon Khan

PIDE2018FPHDETS04

Plagiarism Undertaking

I solemnly declare that research work presented in the thesis titled **"Evaluating The Performance of Variable Selection and Forecasting Methods Using Big Data: A Monte Carlo Simulation Approach"** is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the **HEC** and **Pakistan Institute of Development Economics, Islamabad** towards plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of PhD degree, the University reserves the rights to withdraw/revoke my PhD degree and that HEC and the University has the right to publish my name on the HEC/University Website on which names of students are placed who submitted plagiarized thesis.

Students/Author Signature: Faridoon

Mr. Faridoon Khan

PIDE2018FPHDETS04

Certificate of Approval

This is to certify that the research work presented in this thesis, entitled: **"Evaluating The Performance of Variable Selection and Forecasting Methods Using Big Data: A Monte Carlo Simulation Approach"** was conducted by **Mr. Faridoon Khan** under the supervision of **Dr. Amena Urooj and Co-Supervisor Dr. Saud Ahmed Khan** No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Econometrics from **Pakistan Institute of Development Economics, Islamabad.**

Student Name: Mr. Faridoon Khan
PIDE2018FPHDETS04
Signature: _____

Examination Committee:

a) **External Examiner:**
Dr. Eatzaz Ahmed
Professor
School of Economics QAU, Islamabad
Signature: _____

b) **Internal Examiner:**
Dr. Ahsan ul Haq
Assistant Professor,
PIDE, Islamabad
Signature: _____

Supervisor:
Dr. Amena Urooj
Assistant Professor PIDE,
Islamabad
Signature: _____

Co-Supervisor
Dr. Saud Ahmed Khan
Assistant Professor PIDE,
Islamabad
Signature: _____

Dr. Shujaat Farooq
Head, PIDE School of Economics (PSE)
PIDE, Islamabad
Signature: _____

Abstract

Statistical learning has two primary goals: ensuring high prediction accuracy and discovering relevant predictive variables. Variable selection is crucial when the representation of the true underlying model is sparse. Finding important predictors will improve the fitted model's ability to forecast. Numerous methods for selecting variables are discussed in the literature, but different methods select a different subset of variables and also vary their performance under distinct circumstances. We can evaluate their relative performance by comparing them. This study compares Autometrics and machine learning techniques, including Minimax Concave Penalty (MCP), Elastic Smoothly Clipped Absolute Deviation (E-SCAD), and Adaptive Elastic Net (AEnet). For simulation experiments, three kinds of scenarios are considered by allowing multicollinearity, heteroscedasticity, and autocorrelation conditions with varying sample sizes and a varied number of covariates. First, we evaluate the performance under huge big data. In the presence of low and moderate cases of multicollinearity and autocorrelation, the considered methods retain all relevant variables, but MCP and E-SCAD over-specify the true data generating process (DGP). In the presence of extreme multicollinearity and Autocorrelation cases, the AEnet showed better performance comparatively. In case of heteroscedasticity, the AEnet specifies the true DGP very efficiently. Similarly, the forecasting performance of these methods, including factor models, is evaluated under the same conditions. The MCP produced more accurate forecasts than the rival methods, excluding a few cases where the proposed factor model and E-SCAD outperformed the competitors. While considering the fat big data, E-SCAD remains very effective in contrast to competing approaches in terms of variable selection. Under the forecasting exercises, the Autometrics remained quite successful. Complementing the simulation exercise, we have carried out an empirical application on a popular macroeconomic and financial dataset in Pakistan. The empirical results supported the results of the simulation experiments.

Acknowledgements

First and foremost, I am thankful to Almighty Allah, the most gracious, the most merciful and the magnificent, who bestowed me with the abilities and skills to complete my PhD. I bow in front of him in gratitude for all His blessings. I pay my heartiest tributes to the Holy Prophet Muhammad (Peace Be Upon Him) who is forever the source of guidance and knowledge for the humanity as a whole.

I would like to express my gratitude to my Supervisor, Dr. Amena Urooj and Co-Supervisor Dr. Saud Ahmed Khan, for their supervision, support, and encouragement throughout this research work Without their precious support it would not be possible to conduct this research. In fact, I am privilege to carry out my research under the supervision of such outstanding professors.

I am also grateful to Dr. Eatzaz Ahmad for reviewing my thesis and providing very constructive comments and suggestions for improving my research work. I am benefited and steered from the constructive comments, criticism and suggestions of Dr. Ahsan-ul-Haq and Dr. Zahid Asghar. I would also like to express my gratitude to all my respectable teachers: Dr. Atiq ur Rehman, Dr. Hafsa Hina, and Dr. Uzma Zia for their nice attitudes and setting themselves as examples of great teachers.

I am indebted to my colleagues and friends who helped me throughout this tenure. Especially my discussions with Dr. Sara and Dr. Imran Khan were very important regarding the simulation study and shrinkage methods. I am lucky to have enjoyed the company of Ali Asghar, Asad Shahbaz, Tariq Majeed, Rizwan Ahmad and Nauman Ahmad. Special thanks go to Saira Rasul, Hifsa Mobeen, Rameez Malik, Salman Shah, Javeria Sarwar and Farah Naz, who always encouraged me during my PhD.

I express deep admiration for my Mother and Father who has been a permanent source of love, hope, guidance and kindness for me right from the beginning of my life. Without their assistance, I would not have been able to complete my PhD. I would like to thanks my brothers and sisters, who have consistently prayed for my success and helped me throughout my PhD. When I had free time, my nephews Mueed Anwar and Zaryab Anwar kept me entertained.

TABLE OF CONTENTS

Chapter 1.....1
Introduction.....1
1.1. Big Data and its Significance.....1
1.2. Tools for Big Data.....3
1.3. Objectives of the Study.....8
1.4. Significance of the Study.....8
1.5. Organization of the Thesis.....9
Chapter 2.....11
Literature Review.....11
2.1. Literature Review Related to Variable Selection Methods.....12
2.1.1. Shrinkage Methods.....12
2.1.2. General Unrestricted Model (GUM) and Autometrics.....17
2.2.1. Fresh Insights from Experiments and Implications.....20
2.3. Literature on Remittances, Stock Markets, and Inflation.....25
2.3.1. Studies Related to Workers' Remittance.....25
2.3.2. Studies Related to Inflation.....28
2.3.3. Studies Related to Stock Markets.....31
2.4. Rationale of the Study.....33
Chapter 3.....35
Methodology.....35
3.1. Autometrics.....36
3.1.1. Methodology.....37
3.2. Shrinkage Methods.....39
3.2.1. Elastic Smoothly Clipped Absolute Deviation (ESCAD).....39
3.2.2. Minimax Concave Penalty.....41
3.2.3. Adaptive Elastic net (AEnet).....41

3.3. Factor Models 43
3.3.1. Principal Component Regression (PCR) 43
3.3.2. The Partial Least Squares (PLS) Method 44
3.4. Selection of Tuning Parameter(s) 45
3.5. Simulation Study 46
3.5.1. Data Generating Process 47
3.5.2. Measures of Methods Performance 50
Chapter 4 53
Results and Discussion 53
4.1. Comparison of Feature Selection Procedures using Huge Big Data 53
4.1.1. Design of Experiments and Results 54
4.1.2. SCENARIO-I 55
4.1.3. SCENARIO-II 59
4.1.4. SCENARIO-III 61
4.2. Comparison of Feature Selection Procedures using Fat Big Data 64
4.2.1. Design of Experiments and Results 64
4.2.2. SCENARIO-I 65
4.2.3. SCENARIO-II 70
4.2.4. SCENARIO-III 74
Chapter 5 77
Forecast Comparison and Discussion 77
5.1. Out-of-Sample Forecasting Comparison using Huge Big Data 77
5.1.1. Simulation Results 77
5.2. Out-of-Sample Forecasting Comparison using Fat Big Data 87
5.2.1. Simulation Results 88
Chapter 6 101
Real Data Implications 101

6.1. Comparison of Variable Selection Methods using Huge Big Data 101
6.1.2. Correlation matrix 102
6.2. Comparison of Variable Selection Methods using Fat Big Data 106
6.3. Out-of-Sample Forecasting Comparison using Huge Big Data 109
6.3.1. Data Source 109
6.3.2. Correlation Matrix 110
6.3.3. Forecast Comparison Based on Dual Real Datasets 113
6.4. Out-of-Sample Forecasting Comparison using Fat Big Data 115
6.4.1. Inflation Forecasting 116
Chapter 7 118
Conclusion, Limitations and Future Direction 118
7.1. Limitations and Future Direction 121
References 123
Appendix A 139
Appendix B 148
Appendix C 153

LIST OF FIGURES

Figure 1.1 Types of Big Data....2

Figure 1.2 Advanced Statistical and Penalized regression methods....3

Figure 3.1 Major points of the Autometrics approach....38

Figure 4.1 Distribution of candidate variables into relevant (p) and irrelevant (q) 54

Figure 4.2 Potency under low and high cases of multicollinearity, when n = 80 and P = 50(a) and P = 70(b). 58

Figure 4.3 Potency and Gauge across when ρ = 0.90, n = 80 and P = 50(a) and P = 70(b)....63

Figure 4.4 Distribution of candidate variables into relevant and irrelevant variables 65

Figure 4.5 Computation of Potency across sample sizes when $\sum$ = 0.25, P = 130(a) and P = 150(b) 68

Figure 4.6: Computation of Potency and gauge across low, moderate, and high levels of multicollinearity when n = 40 and P = 130(a), P=150(b) 69

Figure 4.7: Computation of Potency across sample sizes when $\boldsymbol{\pi_1}$ = 0.1/0.3, P = 130 72

Figure 4.8: Computation of Potency across all levels of Heteroscedasticity when n = 40 and P = 130....72

Figure 4.9: Computation of Potency across all levels of Heteroscedasticity when n = 40, P = 130 and P=150 73

Figure 4.10: Computation of Potency across the sample sizes when ρ = 0.25, P = 130. 76

Figure 4.11: Computation of Potency across all levels of Autocorrelation when n = 40, P = 130 76

Figure 5.1: Out of sample root mean squares error across sample sizes, where forecasts are obtained from various models when ρ = 0.25(a) and P = 50 80

Figure 5.2: Out of sample root mean squares error across the levels of Multicollinearity, where forecasts are obtained from various models when n = 80(a), n = 320(b), and P = 70 81

Figure 5.3: Out of sample root mean squares error across sample size, where forecasts are computed from different models when rho = 025(a), 0.5(b), 0.9(c) and P = 70 86

Figure 5.4: Out of sample root mean squares error across the levels of Autocorrelation, where forecasts are obtained from various models when n = 80(a), 320(b) and P = 70.....87

Figure 5.5: Out of sample root mean squares error across sample sizes, where forecasts are computed from various models when rho = 0.25, rho = 0.5, rho = 0.90 and P = 130 92

Figure 5.6: Out of sample root mean squares error across sample sizes, where forecasts are computed from several models when rho = 0.9, P = 130(a) and P =150(b) 93

Figure 5.7: Out of sample root mean squares error across sample sizes, where forecasts are computed from several models when π_1 = 0.1/0.3(a), π_2 = 0.2/0.6(b) and P =130 95

Figure 5.8: Out of sample root mean squares error across sample sizes, where forecasts are computed from several models when π_1 = 0.3/0.9, P =130(a) and P =130(b) 96

Figure 5.9: Out of sample root mean squares error across sample sizes, where forecasts are computed from several models when ρ = 0.25(a), ρ = 0.50(b), ρ = 0.90(c) and P =130 100

Figure 5.10: Out of sample root mean squares error across the levels of Autocorrelation, where forecasts are computed from several models when n = 40 and P =150.....100

Figure 6.1: Correlation structure among covariates 103

Figure 6.2: Trend of Workers' Remittance series (in the log) against time 109

Figure 6.3: Trend of Stock Prices series (in the log) against time 110

Figure 6.4: Pairwise correlation using the determinants of workers' remittance 111

Figure 6.5: Density Plot of Pakistan Stock Prices 113

Figure 6.6: The proposed model versus the baseline models (Workers' Remittance series) 114

Figure 6.7: The proposed model versus the baseline models (Stock prices series).....115

Figure 6.8: Monthly inflation detrended series against time 116

Figure 6.9: The proposed model versus the baseline models (Inflation series).....117

LIST OF TABLES

Table 4.1 Variable Selection under Multicollinearity from Monte Carlo Simulation 57

Table 4.2 Variable Selection under Heteroscedasticity from Monte Carlo Simulation 60

Table 4.3 Variable Selection under Autocorrelation from Monte Carlo Simulation 63

Table 4.4 Variable Selection under Multicollinearity from Monte Carlo Simulation 68

Table 4.5 Variable selection under Heteroscedasticity from Monte Carlo Simulation 72

Table 4.6 Variable selection under Autocorrelation from Monte Carlo Simulation 76

Table 5.1 Forecast Comparison under Multicollinearity from Monte Carlo Simulation 80

Table 5.2. Forecast Comparison under Heteroscedasticity from Monte Carlo Simulation 83

Table 5.3. Forecast comparison under Autocorrelation from Monte Carlo Simulation 85

Table 5.4. Forecast comparison under Multicollinearity from Monte Carlo Simulation 91

Table 5.5. Forecast comparison under Heteroscedasticity from Monte Carlo Simulation 95

Table 5.6. Forecast comparison under Autocorrelation from Monte Carlo Simulation 99

Table 6.1. Features Selection based on Real Data (Huge Big data) 105

Table 6.2. Features Selection based on Real Data (Huge Big data) 107

Table 6.3. Features Selection based on Real Data (Fat Big data).....109

Table 6.4. Pairwise Correlation using the Determinants of Stock Market 113

LIST OF ACRONYMS

DGP	Data Generating Process
LDGP	Local Data Generating Process
OLS	Ordinary Least Squares
GUM	Generalized Unrestricted Model
MCP	Minimax Concave Penalty
SCAD	Smoothly Clipped Absolute Deviation
ESCAD	Elastic Smoothly Clipped Absolute Deviation
LASSO	Least Absolute Subset Selection Operator
EN	Elastic Net
AEnet	Adaptive Elastic Net
ALasso	Adaptive Lasso
Gets	General to Specific
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
FM	Factor Models
DFA	Dynamic Factor Models
PCA	Principal Component Analysis
PLS	Partial Least Square
DI	Diffusion Index
CV	Cross Validation
ML	Machine Learning
WDI	World Development Indicators
IFS	International Financial Statistics
ICRG	International Country Risk Guide

SBP State Bank of Pakistan

INF Inflation

REM Remittance

BLUE Best Linear Unbiased Estimators

LUE Linear Unbiased Estimators

Chapter 1

Introduction

Regression analysis is a well known statistical approach that is used in a wide range of areas, from finance to the social sciences. The prime focus of regression analysis to model the impact of one or more regressors on response variable. The ordinary least squares (OLS) method is commonly used to estimate the unknown parameters of a regression model (Filzmoser and Nordhausen, 2021). The estimates of OLS are obtained by minimizing the residuals squared errors. It is very popular approach because it is easily interpretable and produces best estimates if the underlying assumptions are satistfied (Gujarati et al., 2012).

In the era of big data, the format of data sets has evolved. In the past, the number of observations, n, often much larger than the number of explanatory variables, p, but nowadays, $n \approx p$ or even $n < p$ is common, which is referred to as high-dimensional data. This kind of vast data sets offered new challenges including degrees of freedom, multicollinearity, heteroscdasticity etc, which make the classical linear regression models inefficient. In other words, traditional econometric models do not yield sparse models and therefore may exhibit inefficient behavior when $n < p$. Advanced regression techniques are therefore required for large data sets, also refers to big data (Kim and Swanson, 2013).

1.1. Big Data and its Significance

The era of big data presents an appealing prospect for new economic and econometric advancements. In economics research, the amount of data observed and used in practice is expanding at a rapid rate, and it is widely acknowledged that big data has the potential to significantly impact both economic study and economic policy (Eisenstein and Lodish, 2002).

Thus, economic data is important capital that can be used to make decisions about the economy and the social and economic state of society (competitive intelligence, strategic intelligence, and others), with a particular emphasis on big data (Robles et al., 2019). However, the term "big data" has multiple definitions. In the context of regression, big data was categorized into three types by Doornik and Hendry (2015): tall big data; huge big data; and fat big data. Each category can be described as:

- Tall big data: more observations and several covariates
- Huge big data: more observations and more covariates (observations exceed covariates).
- Fat Big data: Fever observations and more covariates.

Figure 1.1 is a graphical representation of the Big Data kinds. Now the question is, how to cope such Big data. In the next subsection, the suitable methodologies are briefly discussed.

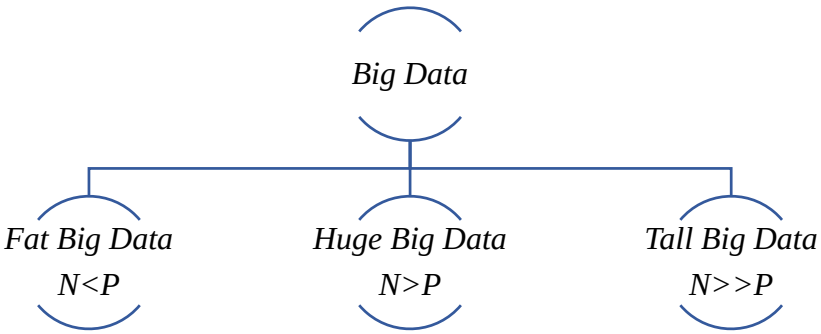


***Figure 1.1*:** Types of Big Data

Numerous fields, including environmental economics, monetary policy research, and macroeconomics, rely heavily on accurate predictions of macroeconomic variables. An increase in the forecasting accuracy lead to a greater understanding of economic dynamics (Bai and Ng,

2008), more effective monetary policies (Bernanke et al. 2005), and improved portfolio management and hedging methods (Rapach et al. 2010). Many macroeconomic statistics are tracked by economists and decision-makers in today's data-rich environment.

1.2. Tools for Big Data

It is supposed that dealing with big data is not an easy task, and to date, there are limited methodologies in the literature that can be used to improve least squares estimates in a data-rich environment. A data-rich environment and big data are used interchangeably in our study. All commonly used approaches and their modified versions are depicted in Figure 1.2, which are often highly appropriate to overcome big data.

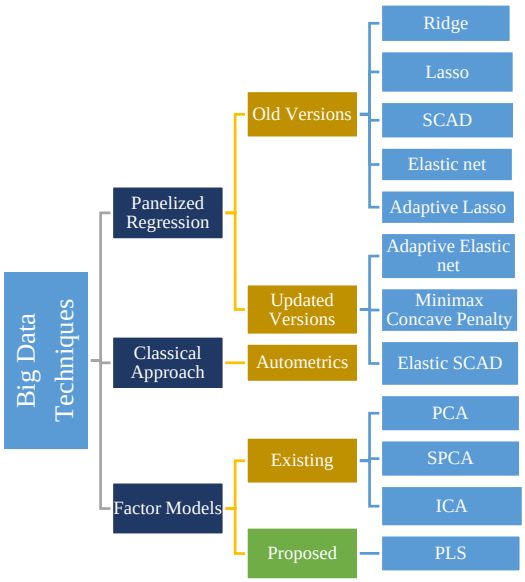


Figure 1.2: Advanced Statistical and Penalized Regression Methods

Panelized least squares algorithms are an essential part of machine learning (ML). It has already been demonstrated in the literature that ML approaches are effective for evaluating large amounts of data (Castle et al. 2021). In general, the penalized regression methods are modified versions of

ordinary least squares regression (OLS). The modified form of OLS can be written mathematically as follows:

$$\sum_{c=1}^{n}(y_c - \alpha_0 - \sum_{d=1}^{m}\alpha_d x_{cd})^2 + k * \vartheta \sum_{d=1}^{m}|\alpha_d| + k * (1-\vartheta)\sum_{d=1}^{m}\alpha_d^2 \quad (1.1)$$

In Equ. (1.1), y_c and x_{cd} are the output and input variables respectively, α_d (*d=0, 1, 2, ..., m*) indicates the unknown parameters to be estimated from the data at hand. Like in classical regression, the first component is the sum of squared residuals and the remaining part represents the shrinkage penalty. Here 'k' refers to the tuning parameter and is often selected by cross-validation. The other parameter is ϑ, and by altering its value, we get different models. More specifically, equating ϑ = 0, results in the ridge regression model, if ϑ = 1 is taken as there is in Lasso regression, and for the value of ϑ between zero and one, we get the model for elastic net (James et al., 2013). As their name suggests, penalized least square methods are based on some constraints or penalties. A good penalty consists of the following three oracle properties: unbiasedness, continuity, and sparsity (Algamal and Lee, 2015). Several methods belonging to the family of penalized regression like Ridge, Lasso, and Elastic Net do not satisfy all the aforementioned oracle properties (Fan and Li, 2001; Zou, 2006). Although in literature, some modified methods from the family of penalized regression satisfy the required oracle properties including Smoothly Clipped Absolute Deviation (SCAD) and Adaptive Least Absolute Shrinkage and Selection Operator (ALasso). Despite satisfying the oracle properties, both the SCAD and ALasso selects only one variable from a group of correlated covariates and ignores other variables. The selected variables may or may not be theoretically important. The SCAD was modified by adding another property to its penalty, which spurs a set of highly correlated covariates to be in or out of the model at the same time. In other words, the new version of SCAD is capable of selecting a group of correlated variables instead of a single one. Similarly, the elastic net is modified in the

form of an adaptive elastic net (AEnet), which achieves an oracle property. The AEnet is capable of including or excluding a set of features simultaneously. In 2010, Zhang developed another penalty model known as the Minimax Concave Penalty (MCP). The method also enjoys an oracle property. To summarize the whole discussion related to penalized regression tools, adaptive elastic net, MCP, and elastic SCAD are the updated forms, primarily used for variable selection and forecasting, and will be elaborately explored in the next sections.

Another approach for automatic model selection was proposed by (Hoover et al., 1999; Krolzig and Hendry, 2001), known as PcGets. This method is based on the idea of general to specific **(**gets) modeling. It starts with a general unrestricted model that captures the key attributes of the underlying dataset. Their standard testing approaches are utilized to decrease its complexity by removing statistically insignificant variables and inspecting the validity of the reductions at every stage to ensure the congruence of the selected model. They studied PcGets' probabilities of recovering the data generating process (DGP) through Monte Carlo experiments and got reliable results. The consistency of the PcGets procedure was established by (Campos et al. 2003).

The new version of the PcGets algorithm was proposed by Doornik (2009b), as Autometrics. This version is based on the same principles as PcGets. Autometrics utilizes a tree-path search to identify and knock out statistically insignificant covariates. Although if the relevant covariate is eliminated by chance, the algorithm works and does not get stuck even in a single route, containing other covariates as proxies (like in stepwise regression). The beauty of this algorithm is that it works well even if the number of covariates exceeds the number of observations (Castle et al. 2021).

In the literature, several simulation studies have investigated the performance of variable selection procedures in the presence of multicollinearity (moderate and severe cases) and high variance of the error term, but, on the application side, they often apply them to cross-sectional data. For instance, SCAD was assessed against several tools, including Lasso and ridge regression, in case of moderate multicollinearity. It was noted that Lasso produces good results when the sample size is small and the variance is high. As the variance was reduced, SCAD performance improved, as shown by (Fan and Li, 2001). Increasing the sample and decreasing the variance (moderate level of multicollinearity) makes the problem of feature selection easier for adaptive Lasso. It was shown by Zou (2006). Similarly, in cases of large sample size and moderate correlation among the explanatory variables, SCAD and adaptive elastic net provide outstanding results. Increasing the correlation level among the explanatory variables adversely influences the performance of SCAD, as shown by (Zou and Zhang, 2009). They also showed, using different levels of multicollinearity and sample sizes, that the adaptive elastic net is more robust than Lasso and adaptive Lasso. Zeng and Xie (2014) revealed that errors increase with increasing variance. They showed that group feature selection procedures are more powerful than single feature selection procedures. In the severe case of multicollinearity, the elastic net is more robust in contrast to ridge regression and Lasso (Zou and Hastie, 2005). Muhammadullah et al. (2022) evaluated the performance of weighted lag adaptive Lasso (WLALasso) with Autometrics in terms of feature selection and forecasting. The simulation experiments demonstrate that in presence of strong linear dependency amidst covariates, the WLALasso outperformed the Autometrics and adaptive Lasso.

When forecasting financial or economic variables, it is often essential to include a piece of information from a large set of potential independent variables in the forecasting model. It is more probable to face the problem of multicollinearity in presence of such a huge set of covariates, and

the construction of factor models circumvents the problem of multicollinearity without losing important information. Most traditional macro-econometric prediction techniques, albeit, are incapable dealing with this, either because it is inefficient or impossible to include a large number of features in a single forecasting model and estimate them utilizing standard statistical tools. One of the alternatives to this problem is factor-based regression models, which have gained prominence. An influential application in Stock and Watson (2002b), is where a limited number of principal components are extracted from a large data set and added to a standard linear regression model, which is utilized to forecast key macroeconomic variables. Stock and Watson (2002a) and Bai (2003b) came up with the asymptotic theory that makes it possible to use principal components to find common factors in large data sets.

In econometrics, researchers has spent considerable effort on developing tests and selection criteria to discover the number of factors that delineate the best dynamics in a massive set of predictors. A paramount contribution was made by Bai and Ng (2002), who developed a range of consistent information criteria that can be used to determine the common factor space underlying a large panel of covariates. The number of factors that are selected in such a way yields an upper bound for the number of factors that should enter the forecasting model for a certain variable. There is no theoretical ground to allow all factors to enter into the forecasting regression. Hence, it is of great importance that a form of factor selection is carried out that is tailored to determining a factor-based forecasting model for a specific variable.

Recent studies have shown that factor models (FM) may provide a parsimonious way to include incoming information about a wide variety of economic activities (Gavin and Kliesen, 2006). These models use a large dataset to extract a few common factors. Many researchers, including

Stock and Watson (1999, 2002a), Bernanke and Boivin (2003), Bernanke et al., (2005), Giannone et al., (2005), Bai and Ng (2006b, 2008), and Castle et al., (2013), have promoted the idea that factor models can be used to enhance macro-econometric models' predictive ability. Diffusion indices and factor models are now quite widely used for economic forecasting (Forni et al., 2000; Peña and Poncela, 2004; Schumacher and Breitung, 2008).

1.3. Objectives of the Study

First objective seeks to find the best variable selection technique under a variety of simulated scenarios including sample size, candidate variables, multicollinearity, heteroscedasticity and autocorrelation. The simulated scenarios are altered in order to thoroughly investigate the procedures under consideration. The second objective consist of proposing the novel factor model: its forecasting performance is compared with existing models. The comparison is made under different scenarios including sample size, candidate variables, multicollinearity, heteroscedasticity and autocorrelation. Different simulated scenarios are considered to examine the techniques in depth. Complementing the simulation exercises, the performance of variable selection techniques is evaluated using macroeconomic and financial datasets in third objective. In the fourth objective, the performance of proposed factor model is evaluated against existing techniques using macroeconomic and financial datasets.

1.4. Significance of the Study

In recent years, econometricians and statisticians have paid a significant amount of attention to the big data environment, variable selection, and data mining methodologies. Increasing independent variables present a number of problems for the econometric model, such as multicollinearity, degrees of freedom, and heteroscedasticity. Consequently, most traditional time series models, such as vector auto-regression (VAR) and vector error correction model (VECM) do not perform

well. These methods adjust no more than ten covariates, as more covariates generate major problems that render the results invalid (Stock and Watson, 2002). In contrast, macroeconomic variables such as inflation, economic growth, and remittances inflow are derived from a large number of candidate features (covariates). Now the question is raised here: which subset of features statistically influences the response variable? This is a challenging task for all feature selection techniques.

This work employs several advanced statistical and machine learning techniques to address the aforementioned issues and produce reliable results. The techniques will be compared under simulated scenarios for multicollinearity, heteroscedasticity, and autocorrelation before being applied to macroeconomic and financial datasets to provide conclusive answers regarding the predictability and validity of distinct theoretical scenarios simultaneously. The key goal of our research is to come up with a better way to help policymakers. This better tool can be used with any macroeconomic and financial time series datasets in a rich data environment, not just workers' remittances, stock market, and inflation data.

1.5. Organization of the Thesis

The remaining part of the thesis is organized as follows.

In section 2, we discuss past studies regarding theoretical and empirical aspects of variable selection and predictive modeling tools.

In Section 3, we discuss the big data techniques including the classical approach (Autometrics), shrinkage methods, and factor models. In addition, we discuss the data generating process under different simulation scenarios like multicollinearity, heteroscedasticity and autocorrelation.

Section 4 deals with the variable selection procedures, in which we have two main subsections. In both subsections, Monte Carlo evidence on the comparative performance of several variable selection techniques is discussed separately.

Section 5 deals with the forecasting comparison, in which we have two main subsections. In both subsections, we evaluate the predictive power of the proposed factor model against existing techniques through simulation exercises.

Section 6 deals with real data analysis, in which there are four key subsections. The first two subsections cover the comparison of variable selection techniques, and the last two subsections explore the predictive power of the proposed factor model against existing techniques.

Section 7 provides a detailed discussion of simulation findings. Also, it delineates the real application of our study. The limitations and future directions for research are also given in this section.

Chapter 2

Literature Review

Regression is a fundamental statistical component that is most commonly used to construct a prediction model. In regression analysis, the goal is to create a model that is accurate in both the selection of important features and prediction. Various regression approaches are used to develop models in numerous domains, including biology, business, and other social sciences. A traditional linear regression model is one of them, and it is most often used to build a model linearly due to its ease of mathematical calculation and simple understanding. However, there are various scenarios in which the typical linear regression model does not perform well and sometimes even produces inaccurate results, particularly when dealing with large amounts of data (more covariates). Consider the following issues in a rich data environment:

i) Multicollinearity

ii) Degrees of freedom

iii) Large variability

A list of methods has been suggested by researchers in past studies to handle these issues. Literature on this topic started with stepwise regression (Breaux, 1967) and Ridge regression (Hoerl & Kennard, 1970). Moving to more advanced methods, non-negative garrote (Breiman, 1995), the least absolute shrinkage and selection operator (Tibshirani, 1996), Elastic net (Zou & Hastie, 2005), Adaptive Lasso (Zou, 2006), Adaptive elastic net (Zou and Zhang, 2009), Smoothly Clipped Absolute deviation (Fan and Li, 2001), Autometrics (Doornik, 2009b), Minimax Concave penalty (Zhang, 2010), Elastic SCAD (Zeng and Xie, 2014) and factor models (Stock and Watson, 2002a).

A literature review is divided into three subsections. Section 2.1 provides a comprehensive review of the theoretical and empirical implications of variable selection methods, followed by forecasting tools in Section 2.2. Section 2.3 elaborately delineates the past studies related to inflation, workers' remittance and the stock market, and the last section demonstrates the rationale of the study.

2.1. Literature Review Related to Variable Selection Methods

Feature selection has become a crucial area in time series analysis in the recent era. In general, a huge set of covariates is often utilized at the initial stage of analysis to mitigate possible modelling biases. Here, we discuss the past studies related to two different approaches (i.e., shrinkage methods and Autometrics) to variable selection.

2.1.1. Shrinkage Methods

The literature on subset selection started with stepwise regression (Breaux, 1967). The subset selection procedure is a discrete process, where either the variable is dropped or retained according to statistical significance. It produces an interpretable model but can be highly variable. The minor variations in the data induce considerably different results to be selected, and this can attenuate their predictive power.

The subset selection procedure yields an interpretable model but can be extremely variable because it is a discrete process, which means, that either the variable is dropped or retained in the underlying model. Small changes in data result in very different models being selected, and this can vary their forecasting accuracy. The subset selection methods suffer from high variability (Tibshirani, 1996), and neglect the stochastic error when selecting the subset from a large set of variables. Thus, it is somewhat hard to understand their theoretical properties (Fan and Li, 2001).

Ridge regression was developed by Hoerl and Kennard (1970), which imposes L_2 penalty and is more stable due to forcing the coefficients towards zero. Unlike the advanced variable selection approaches, it tends the variables towards zero exactly, and thereby, the model is not easily interpretable (Zou and Hastie, 2005; James et al., 2013).

Breiman (1995) performed a simulation experiment in which the proposed method, which is non-negative garrote (nn-garrote), was compared with existing methods, namely subset selection and ridge regression. According to findings, the method of subset selection is not stable, but in contrast, the ridge regression ensures stability, and the nn-garrote is moderately stable. Consequently, the prediction accuracy of nn-garrote is reduced compared to ridge regression due to the slight instability it produces. He showed that in variable selection, the nn-garrote method usually beats rival tools, including ridge regression and subset regression.

Cantoni et al., (2009) extended a model selection procedure based on nn-garrote in a non-parametric way and compared it with other rival procedures. The predictive power and correct variable selection have remained good for the proposed approach. Similarly, another feature was included in the nn-garrote method to make it robust against leverage points and vertical outliers. The proposed approach is assessed through a simulation exercise in comparison with several other approaches. The proposed version often beats the existing tools, which are also supported by applications on real data (Gijbels and Vrinssen, 2014).

A promising procedure is “so-called” the least absolute shrinkage and selection operator (Lasso) was first proposed by Tibshirani (1996). Basically, the Lasso belongs to the regularized least-squares family and imposes an L_1-penalty on the estimated coefficients. Due to its L_1-penalty nature, the Lasso performs both automatic variable selection and continuous shrinkage simultaneously. Owing to this, Lasso produced the output in such a way that several coefficients

are exactly zero (sparsity) and hence provide an interpretable model. In general, a penalty term is considered good if it satisfies the oracle properties, namely sparsity, unbiasedness, and continuity. The properties of the good penalty function proposed by (Fan and Li, 2001) are listed below.

- The resultant estimator promotes sparsity by setting redundant variables to zero in order to simplify the model.
- Unbiasedness: When the true parameter is unknown, the resultant estimator is assumed to be unbiased.
- Continuity: In order to reduce the instability in model prediction, the resultant estimator must be continuous.

Through simulation and real data experiments, Tibshirani (1996) compared Lasso with subset selection and ridge regression and found that Lasso is considerably better than the other two procedures. Fan and Li (2001) argued that the Lasso estimator is inconsistent, in that it penalizes all the estimated coefficients uniformly, and consequently, irrelevant variables are over-penalized in the form of biased estimators.

One of the main reasons for the Lasso not being consistent, i.e., lacking the oracle property (Fan and Li, 2001) is that it equally penalizes all the coefficients, which over-penalizes the irrelevant variables, leading it to be a biased estimator. Related to this drawback of Lasso, a new version of Lasso was proposed by Zou (2006), named Adaptive Lasso (ALasso). The ALasso penalizes different coefficients in the L_1-norm using adaptive weights, with small weights assigned to large estimates and vice versa. Consequently, the selection bias tends to zero, and we attain unbiased and consistent estimates. Apart from this, the ALasso solution is continuous as well, which enables

it to fulfil the oracle properties. As shown by Zou, ALasso is more robust in terms of variable selection in contrast to Lasso, Ridge, and a few other techniques.

The L_1-penalty methods can have very poor performance when the correlation among the covariates set is sufficiently large (Zou and Zhang, 2009). The problem of multicollinearity is more likely to arise in the case of high-dimensional data analysis. More specifically, if the predictors are highly correlated to each other, then the performance of Lasso considerably deteriorates, and its path is also unstable, as shown in Zou and Hastie (2005). On the other hand, if there is no association among the features but the features' dimension is high, the maximum sample association could be large (Fan and Lv, 2008).

The penalties L_1 and L_2 were combined, and we called it elastic net (Enet). Owing to L_1-norm and L_2-norm, Enet performs automatic variable selection and stabilizes the solution paths, respectively, and thus enhances the forecasting accuracy (Zou and Hastie, 2005). In presence of orthogonal design, the Lasso yields efficient output, as shown by Donoho et al. (1995), and the performance of Enet tends to the Lasso. The high correlation among the features significantly improves the forecasting performance of the Enet against Lasso. They carried out some simulation experiments under different scenarios and found that the elastic net is more robust in terms of variable selection than individual penalties.

From the aforementioned discussion, we can conclude that Enet and adaptive Lasso boost the lasso's performance in two distinct ways. The Enet overcomes the collinearity and adaptive Lasso gains the oracle properties.

Although, following the arguments in (Zou and Hastie, 2005; Zou, 2006), it can be observed that the adaptive Lasso suffers from instability under high-dimensional data and the Enet does not

achieve an oracle property. Thus, to handle the last dual limitation, Fan and Zhang (2009) proposed an adaptive elastic net (AEnet) by combining the two tools, adaptive Lasso and Enet. Through a simulation study, they showed that the AEnet method solves the problem of collinearity better than other methods, leading to better performance for finite samples.

A new method was proposed for feature selection known as smoothly clipped absolute deviation (SCAD), which selects a set of correct features and estimates their coefficients simultaneously and thus providing the confidence intervals for the estimated coefficients. This approach is differentiated from competitive tools in the following ways: The penalty function is symmetric, non-convex on (0, ∞), and has singularities at origin to yield sparse solutions. Moreover, the penalty function is restricted by a constant to alleviate bias and meet specific conditions to provide a continuous solution. Through the simulation experiments, it was shown that SCAD outperforms the existing methods, including Lasso and ridge regression, in variable selection and reduces the bias significantly. On the other hand, a high correlation among predictors deteriorates its performance (Fan and Li, 2001). Because of this drawback, the SCAD was extended by adding L_2 penalty: a penalty function that spurs a set of highly associated covariates to be in or out of the model simultaneously. By virtue of this property, the new version of SCAD allows one to select a group of correlated variables. The modified form is so-called elastic-SCAD (E-SCAD). Another approach refers to the minimax concave penalty (MCP) proposed by Zhang (2010), which attains an oracle property under some regularity conditions. Recently, fruitful insight has also been achieved via a theoretical analysis of the global solution (Kim and Kwon, 2012). Algamal and Lee (2015) compared the performance of a few penalization techniques using high-dimensional data. They found that an adjusted adaptive elastic net is considerably more consistent in selecting genes than the other three rival methods.

2.1.2. General Unrestricted Model (GUM) and Autometrics

Big data has enough benefits for statistical modeling, but it also has problems, such as mistaking correlations for causes, too many false positives, ignoring sampling biases, and using the wrong tools (Doornik and Hendry, 2015).

In advanced macroeconometrics, data must be analysed in a high-dimensional setting, which is also called "fat big data" (Doornik and Hendry, 2015). This is because the typical Local Data Generating Process (LDGP) is very complicated. In such circumstances, even highly expert data analysts and researchers are not able to overcome all the probable search paths. Luckily, advancements in programming software and computational capacity mean that complex models are no longer a limitation on the choice of modeling strategy. Instead, it is based on the valuable properties of the resulting model. Even though many past studies (Leamer, 1978, 1983; Lovell, 1983; Faust and Whiteman, 1997) say that a general-to-specific (gets) search approach is not a good way to find empirical models, it has been shown to work.

The Gets approach in autometrics utilises a multi-path searching method that can blend contracting and expanding searches, which enables it to adjust more predictors than the data points (Doornik, 2009b). This characteristic of Gets is specifically useful for modeling unit root processes. It is also admirable because it chooses a set of covariates instead of the whole model. This makes it more adaptable and open to new ideas.

The Autometrics were assessed against Lasso under a variety of conditions, which required data-based association utilising huge datasets. In the first experiment, all variables were orthogonal and irrelevant (null model). Autometrics selected the null model accurately. Including irrelevant lags, the same method again specified the model correctly. In the second experiment, when 10 out of 20

variables were relevant, Autometrics predicted a 60% chance of keeping the covariate with the smallest coefficient (alpha = 0.0001). They revealed that difficulty arises when the covariates are associated with each other. In the presence of high correlation (multicollinearity) among features, the static form of Lasso was shown to have better performance in potency but worse in the form of high gauge (Doornik and Hendry, 2015). Later, the Autometrics performance was compared to statistical learning methods such as Lasso and adaptive Lasso over simulation and real-world datasets. Concerning the parameter estimation, Autometrics generated the least average variance and bias, as anticipated by the definition of the least-squares estimation approach when the right features are chosen. Concerning the feature selection, adaptive Lasso does better than Autometrics in most of the simulated schemes.

2.2. Literature Review Related to Macroeconomic Forecasting

There are many related studies on macroeconomic forecasting based on factor models, machine learning techniques, and Autometrics. Under the massive features environment, the factor-based models, constructed from principal component analysis developed by Stock and Watson (1999, 2002) have been applied in numerous applications, including those of Artis et al. (2005); Bai and Ng (2002, 2006); Boivin and Ng (2006); Kim and Swanson (2014a, 2018); Castle et al., (2013).

The PCA approach provides the factors very consistently for the estimation of a DFM model, as shown by (Stock and Watson, 2002b; and Bai and Ng, 2002). As shown by Stock and Watson (2002b) factor models based on PCA outperformed the univariate auto-regressions and vector auto-regressions in simulation forecasting exercises. The large set of predictors is summarized using a few estimated factors constructed by the PCA approach. Recent statistical studies have

shown (Stock and Watson, 1999) that using a large number of covariates can improve forecasts of key macroeconomic variables by a large amount.

This framework allows the inclusion of data at different frequencies, at different vintages, and at different time spans, thereby yielding a specified and statistically rigorous but economical framework for the use of multiple data sets. PCA-based factor model yields a framework that permits us to remain agnostic regarding the structure of an economy by applying huge amounts of information in the formulation of forecasting experiments. This framework allows the inclusion of data at different frequencies, at different vintages, and at different periods. The new framework is economical and statistically rigorous for the utility of multiple datasets (Bernanke and Boivin, 2003). Factor-based forecasts often beat the standard time series methods (Artis et al., 2005).

For further improvement in forecasting, Bai and Ng (2008) used a factor model based on quadratic principal component analysis to capture non-linearity and yield a more accurate prediction than the existing factor model. Armah and Swanson (2010) argued that the use of suitable proxies in the form of observable economic variables for unobserved factors can be utilized as an alternative to factors in the construction of diffusion index for forecasting. If observable economic variables are indeed good proxies of the unobserved factors, then these proxies could be used in place of factors in the diffusion index model for prediction.

Factor models are more appropriate for short-horizon forecasting, and their forecasting performance deteriorates with an expanding forecast window. To get more accuracy in forecasting, robustness, intercept correction, or differencing strategies are required in case of location shifts, as shown by (Castle et al., 2013). They also described the dataset generated by observable features or hidden factors that can be discovered from the dataset using variable selection tools, depending

on the situation being studied. Whatever the nature of the model, whether it is based on variables or factors, it can be utilized for prediction. Also, real-world evidence has shown how important it is to make the prediction more stable and robust when the location changes.

To gain more improvements in forecasting accuracy, the factor models are combined with robust machine learning tools, namely ridge regression, boosting, bagging, the non-negative garotte, the elastic net, and least angle regression. The combination of forecasting models (hybrids) often outperforms the benchmark and PCA-based factor models (Kim and Swanson, 2014). Including combining approaches, the predictive power can be enhanced by adopting recursively updating and averaging (Castle et al., 2015). As stated by (Bai and Ng, 2008, 2009; Stock and Watson, 2012) factor models, if combined with regularization methods, lead to an improvement in the forecast. As shown by Stock and Watson (2012) empirically, in the case of multiple series under consideration, the dynamic factor model (DFM) often beats the included shrinkage methods forecast.

2.2.1. Fresh Insights from Experiments and Implications

Kim and Swanson (2018) focused on the inspection of several factor estimation techniques, namely independent component analysis (ICA), principal component analysis (PCA), and sparse principal component analysis (SPCA), in conjunction with predictive models, where hybridization is assessed. Using these factor estimation techniques together with several types of penalization and pure machine learning tools, including elastic net, least angle regression, nonnegative garrote, bagging, and boosting. They came to the conclusion that combining these techniques with penalization and machine learning (ML) techniques makes for good forecasts of macroeconomic variables.

Swanson and Xiong (2018a,b) reported some important points regarding the construction of the diffusion index, as it is crucially dependent on whether the data under consideration is real-time or not. In practice, when they estimate the weights for diffusion indexes using real-time big data, rare big data models are found to be superior in contrast to simpler Dynamic Siegal-Nelson (DNS) models after 2010. In addition, in the presence of highly volatile interest rate regimes, ML and other related statistical tools are preferred, whatever the type of data to be used for model construction (Pederson and Swanson, 2019).

Machine learning algorithms provide better forecasts than benchmark and factor models (Richardson et al., 2019). The combined forecast from the factor-based approaches and ML algorithms, including artificial neural networks (ANNs), outperforms the individual factor-based approaches and machine learning, and the application of ML algorithms to common factors is effective in the composite prediction (Maehashi and Shintani, 2020).

Smeekes and Wijler (2018) showed theoretically that lasso type estimators are more robust in forecasting than factor type models and argued that factor type models are not always well suited for macroeconomic forecasting. Using the Lasso technique can lead to a more parsimonious factor model, which can yield better prediction compared to the traditional PCA approach (Kristensen, 2015).

Diebold and Shin (2019) proposed and investigated direct subset-averaging methods based on the partially-egalitarian Lasso structure. After analysis, they learned from the study that, unlike Lasso, the new methods do not require the choice of a hyperparameter. The proposed method beats the simple average as well as median forecasts. Kim and Co (2020) employed partial least square (PLS) on real data along with PCA for factor extraction from a large data set and showed that PLS-

based forecasting models outperformed the benchmark models. As shown by Epprecht et al., (2019), Lasso and adaptive Lasso are more robust in forecasting as compared to Autometrics. Syed and Lee (2021) conducted a time series experiment, in which the dynamic factor model (DFM) was compared with Bayesian machine learning tools and other benchmark models in forecasting the core macroeconomic variables. They concluded that Bayesian machine learning tools are more robust compared to others. As Li and Chen (2014) encountered some interesting findings, forecast combination is deemed another way to boost DFM utilizing penalized estimation. When economic data is noisy, chaotic, and has a lot of dimensions, these highly integrated prediction methods make useful tools for economists.

Castle et al. (2020) assessed the potential of PCA to identify the cointegrating relations in a simulation experiment using a single-equation. In such circumstances, the study explored some issues related to the factor model. In some particular circumstances, the long horizon relationship can be discovered precisely, but the design of the data matrix (both its dimension and correlation structure) influences which PC discovers the long horizon relationship and the precision with which it does so. The PCA approach is unable to identify long-term relations when the variance-covariance matrix is contaminated by irrelevant variables (i.e., big data settings).

Muhammadullah et al. (2022) evaluated the performance of weighted lag adaptive Lasso (WLALasso) with Autometrics in terms of feature selection and forecasting. The simulation experiments demonstrate that in the presence of strong linear dependency amidst covariates, the WLALasso outperformed the Autometrics and regularization tools. Two types of significance levels are considered, which are 1 percent and 5 percent. At a 5 percent significance level, Autometrics retains more irrelevant features, which in turn tends to enhance the root mean square

error relative to the 1 percent significance level. Similarly, Guerard et al. (2020) compared the forecasting performance of Autometrics with some traditional models using macroeconomic series. The study produced a substantive improvement in forecasting accuracy in contrast to traditional models.

As argued by Desboulets (2018), sure independence screening is considered a robust method for linear models, but non-linearities in data distort its performance. Moreover, there is a lack of literature to find a conclusive tool. On the other hand, Khan et al. (2021) argued that a long list of tools exist in the available literature, but unfortunately, there is still no such tool available to the researcher that is dominant in every circumstance. More specifically, a massive body of literature ends up being inconclusive. A substantive development in the past literature can only be gained by devising an ultimate predictive model, which unluckily does not exist. So, we can assume that each procedure works better in certain situations (depending on how the data is generated).

Kock and Terasvirta (2014) conducted an empirical study in which Autometrics is compared with linear and nonlinear models. The tools implied monthly unemployment series and industrial production from the four Scandinavian and G7 countries, and focused on prediction during the economic crisis from 2007 to 2009. The performances of these models are compared by looking at the forecasted models' accuracy. The non-parametric models, namely the artificial network, often outperforms the linear model. Cunha et al. (2019) studied the usefulness of Autometrics and machine learning (ML) algorithms against benchmark models. They failed to gain any improvement in separate Autometrics and ML models against benchmark models. But after weighted combinations, these models achieve substantive improvements in forecasts and beat the benchmark models.

Westerlund et al. (2014) proposed forecast combination (FC) as an alternative to the neural network (NN) and linear regression (LR), the most frequent air quality predictive models. They performed Monte Carlo simulation experiments to evaluate the proposed method against NN and LR. The key findings of the study are that, unless one is lucky enough to select the right model, FC generally beats the LR and NN. The results were supported by the empirical findings. Mansor et al. (2014) performed the simulation exercise and compared the fuzzy approach with the automatic model (Autometrics). They inferred that fuzzy produces accurate results when there are fewer rules and fewer input variables.

Ahumada and Cornejo (2016) analyzed the commodities data to determine, whether considering the cross-dependence improves the forecasting accuracy of individual models. The post-sample forecast period was reasonably unstable, and thereby dual strategies for the forecast were implemented, namely recursive estimations and robust approaches, in order to overcome the adverse impacts of potential breaks. As a result, forecast accuracy was achieved by allowing the interactions of the prices, i.e., joint EqCMs and DVARs. Rocha and Pereira (2019) forecast Brazilian industrial production one step ahead by using Autometrics and the Autoregressive model (AR) as benchmarks. The Autometrics chose the lags based on the saturation of the impulse indicator, which led to a better forecast than the AR model.

Darne and Charles (2020) proposed bridge models to forecast the quarterly gross domestic product (GDP) growth rate for France. The proposed form allows for economic interpretations and is specified by utilizing a statistical approach based on an automatic model selection (general-to-specific approach) and machine learning algorithm, that is, Lasso. These tools are capable of selecting a subset of covariates from a general model. To evaluate their forecasting performance,

a recursive forecast is performed. The bridge models are built through the dual model selection tools and provide better post-sample forecasts against the benchmark models. Finally, the combined forecasts of these tools disclose a remarkable forecasting performance.

Wahid et al. (2017) proposed a Robust Adaptive Lasso (RAL) procedure that utilises the Pearson residuals weighting scheme approach. The weight function assigns fewer weights to such data points, which is inconsistent with the assumed model. It was seen that the RAL estimator retains the relevant variables and produces their estimates as well, in presence of outliers and multicollinearity. They also shed light on the oracle property of model selection and the consistency of the RAL approach. Based on simulation experiments and analysis of real data, it is clear that the proposed method is better than the other shrinkage approaches.

2.3. Literature on Remittances, Stock Markets, and Inflation

This subsection sheds light on the determinants of working remittance, the stock market, and inflation one by one.

2.3.1. Studies Related to Workers' Remittance

Lianos (1997) investigated the influence of family income, migrant income, the rate of inflation, the rate of interest, the rate of unemployment, the exchange rate, and the number of migrants on remittance inflow. The interest rate and the rate of inflation have turned out to be core factors in remittance inflow. El-Sakka and McNabb (1999) implemented the ordinary least squares (OLS) approach to determine the impact of wage, level of domestic income, interest rate, and exchange rate on remittances inflow. They came to the conclusion that interest rates and exchange rates are the most important factors influencing remittance inflows. Arun and Ulka (2010) found that some standard variables like employment, income, and education drive the remittance inflow.

Ahmed et al. (2020) examined a set of variables, including gross domestic product, the transaction cost of sending money, stock of migrants, exchange rate stability, distance, colony, institution, and border on remittances inflow, and found transaction cost to be the core predictor of remittances inflow. Adams (2009) assessed the nexus between skill composition of migrants, poverty, interest and exchange rates and remittances. He found that the skill composition of migrants has a significant impact on remittances. In 2016, Ahmed and Zarzoso found that the remittance flow is adversely affected by transaction costs; revealing that there is the possibility that high costs become a hurdle for migrants and stop sending money to their homes or adopt the informal way of sending money. Silwa and Huang (2005) conducted a study to examine the factors affecting workers' remittances. They found that FFR, unemployment, CPI, money supply are good indicators.

Mustafa and Ali (2018) employed the gravity model for examining the bilateral remittance indicators in case of Pakistan. The list of indicators includes GDP (Home), GDP (source), geographical distance, migrant stock, common official language, colonial linkages, unemployment, and inflation. After analysis, GDP (Home), distance, migration, common official language, and colonial linkages are found to be important indicators for workers' remittances. Similarly, Ahmed and Zarzoso (2014) used GDP (Home), GDP (recipient), geographical remittance, common language, transaction costs, exchange rate, domestic credit to the private sector, interest rate differential, unemployment rate, population density, migrant stock and political stability for bilateral remittances. Excluding the Unemployment rate, exchange rate, and GDP (source), the rest of the variables are driving bilateral remittances. Hina and Ullah (2021) conducted an experiment to explore the determinants of workers’ remittances in case of Pakistan. They used 27 variables as predictors and applied Lasso to discover the important variables that

affected remittance inflow. Following that, the autoregressive distributed lag model was applied to the variables chosen by Lasso, and short and long-horizon associations between remittance inflow and its true determinants were discovered.

Aydas et al. (2005) conducted a study for Turkey in order to trace the determinants of workers' remittances using a list of different predictors. They found that black market premium, interest rate differential, inflation rate, growth, home, and host country income levels, and periods of military administration are significant indicators. Laniran and Adeniyi (2015) conducted a study to find out the factors affecting workers' remittances in Nigeria. They discovered that income per capita, inflation, domestic credit, deposit rate, exchange rate, financial deepening, interest rate differential, secondary school enrollment, and openness all play a significant role in remittances. Ullah et al. (2015) examined the determinants of workers' remittances using VECM in context of Pakistan. The impact of GDP, terrorism index, and trade openness turned out to be significant. A list of variables' influence is evaluated by Abbas et al. (2017), including real gross domestic product, interest rate differential, inflation rate, real effective exchange rate, secondary school enrollment, the number of migrant workers, the financial liberalization index, democracy, internal and external conflicts, D911 in the form of a dummy variable, law and order situation, government stability, corruption, and foreign debts on workers' remittances. Using the generalised method of moments (GMM), it was found that almost all of the predictor variables have a substantial effect on workers' remittances.

Ricketts (2011) investigated the factors that influence worker remittances and concluded that interest rate, inflation, unemployment, exchange rate, and GDP are all good incentives for remittances. Lueth and Ruiz-Arranz (2007) analyzed the nexus between workers' remittances and

GDP, oil prices, the exchange rate, and CPI. They used the VEC model and concluded that the only factors that improve worker remittances are GDP and exchange rate. Gupta (2005) examined the nexus of workers' remittances with drought, LIBOR, change in LIBOR, return on Nasdaq, the Asian crisis, oil prices, S11 dummy, return on Bombay Stock Exchange (BSE), exchange rate change, political uncertainty, geo-political tensions, rating changes, and the issuance of RIB, IMD bonds. After the time series analysis, we arrive at the conclusion that macroeconomic variables considerably promote the workers' remittances.

Akçay and Karasoy (2019) empirically examined the influence of macroeconomic and financial indicators on remittance. They employed the ARDL model and inferred that domestic credit to the private sector, macroeconomic instability, the official exchange rate, macroeconomic instability, the average GDP annual growth rate of OECD countries, and oil prices were related to remittance. Akcay (2021) analyzed the influence of some variables, namely GDP, oil prices, inflation, and foreign direct investment (FDI) on remittances outflows in Saudi Arabia. In the long run, the reaction of remittances outflows to oil prices is asymmetric. Similarly, Abbas (2020) seeks to discover the drivers of remittance. He applied the non-linear panel Pooled Mean Group (PMG) model and came to the conclusion that oil prices, trade openness, FDI, GDP, exchange rate, and financial development significantly contribute to remittances.

2.3.2. Studies Related to Inflation

A study was conducted by Mohanty and John (2015) to establish the relationship between inflation and other predictor variables, namely crude oil prices, output gap, fiscal deficit, and policy rate. The study applied structural vector autoregression (SVAR) to identify the relationships and inferred that inflation dynamics are significantly explained by all the predictors.

Ahmad et al., (2013) seek to estimate the nexus between inflation and several factors using a panel data set including low and high inflation countries. They used the ARDL model and empirical results postulated that imports and GDP are the main drivers of inflation in the case of low inflation countries, while in contrast, national expenditure and money supply are the core factors of inflation.

Shah et al. (2014) assessed the relationship between the consumer price index and a set of covariates such as Producer price index, Gasoline, Imports, Unemployment, Electricity, Employment, Money Supply, Durable Goods, Farm Products, Natural Gas, Steel Mills Product, Crude Petroleum, Oil Production, Agriculture Products Export, Exchange Rate, Capital Goods Import, Food Export, Food import and Government Sector Borrowing for Pakistan. A stepwise regression approach was employed to capture the relevant covariates. It was found that electricity, food import, steel mills product, durable goods, capital goods export, government sector borrowing, and natural gas have an effect on inflation in Pakistan.

Khan and Schimmelpfennig (2006) observed the factors that affect inflation in case of Pakistan. They applied the VECM model and found that monetary variables drive inflation in Pakistan, and are good leading factors for future inflation. Jiranyakul (2019) applied linear and nonlinear cointegration tests along with a two-step procedure and found the long run association between oil prices, industrial production, and inflation.

Qayyum (2006) investigated the relationship between the price level and the factors that determine it. From the analysis, it emerged that the money supply enhances inflation in Pakistan. Similarly, Okoye et al., (2019) evaluated the nexus amid inflation and the major factors causing it in Nigeria. For empirical analysis, advanced statistical techniques were adopted and yielded some interesting

findings, namely that economic growth, external debt, money supply, exchange rate, and fiscal deficits are the key contributors to inflation.

Chiaraah and Nkegbe (2014) evaluated the influence of the exchange rate along with other factors on inflation in Ghana. For this purpose, the error correction model was adopted and concluded that only GDP growth and money supply are the factors that contribute to inflation. Imimole and Enoma (2011) highlighted the exchange rate along with other factors and inflation nexus using the ARDL approach. From the empirical analysis, it is inferred that money supply, exchange rate depreciation, and real gross domestic product are the major factors of inflation in Nigeria. A study was conducted by Mbongo et al. (2014) on inflation in Tanzania. A few time series models are used, and the significant influence of money supply and exchange rate on inflation is discovered.

A study was conducted by Njoku and Nwaimo (2019) to examine the impact of the exchange rate on inflation using the VECM approach. Findings ensured the existence of a long-term relationship between exchange rate and inflation in Nigeria. Furthermore, Non-oil export and money supply fail to drive inflation. Kayamo (2021) seeks to capture the asymmetric association between inflation and exchange rate in Ethiopia by using an advanced tool, referred to the non-linear ARDL model. The empirical findings showed that import and exchange rate are the good determinants of inflation. Moreover, the gross fiscal deficit of the central government, money supply, and real GDP growth have remained ineffective.

Liu and Chen., (2017) assessed the association between import price index, producer's price index, nominal effective exchange rate, money supply, domestic demand, foreign supply and consumer price index in context of China. To trace the cointegration and causality, Johansen and VECM

were applied respectively. As per the study result, the key output is that the exchange rate pass-through a limited but rising influence on domestic prices and will continue to do so.

A study was conducted by Shahbaz (2013) to establish the association between inflation and the factors, namely GDP, money supply, and interest rate. The study applied the ARDL approach and concluded that only the money supply is significantly associated with the inflation rate.

The study was conducted by Jaffri et al. (2014) using time series data in context of Pakistan. They analyzed demand-side factors and supply factors for inflation by implementing the ARDL model. After the analysis, they found that inflation is caused by export and population (demand side) while import, Government Revenue, and Electricity Generation (supply-side).

2.3.3. Studies Related to Stock Markets

A stock exchange market is the centre of a network of transactions where buyers and sellers of securities meet at a specified price. The stock market plays a key role in the mobilization of capital in emerging and developed countries, leading to the growth of industry and commerce in the country, as a consequence of liberalized and globalized policies adopted by most emerging and developed governments. The stock market is one of the most vital components of a free-market economy, as it helps to manage capital for the companies from shareholders in exchange for shares in ownership to the investors. Stock exchanges provide businesses with the facility to raise capital by selling shares to investors (Black and Gilson, 1998).

Shrestha and Subedi (2014) empirically inspected the influence of macroeconomic indicators on the stock market in Nepal. The study applied the OLS approach to discover the main drivers of the stock market. Empirical findings revealed that inflation had been affected by GDP, broad money supply, treasury bill, political events and policy change dummy. Tsaurai (2018) considered a list

of variables that determine the stock market. Empirical findings showed that FDI, Infrastructural Development, Savings, Trade Openness, Exchange Rate, and banks are the key factors for the stock market.

Maku and Atanda (2010) analyzed the impact of the consumer price index, broad money supply, treasury bill rate (a proxy for interest rate), exchange rate, and real output growth on the stock market. The empirical analysis displayed that the Nigerian stock market is more responsive to changes in the inflation rate, exchange rate, real GDP, and money supply. Aamir and Shah (2018) analyzed the factors of stock market co-movement amid Pakistan and Asian emerging economies. The findings of the study demonstrate that there is long-horizon integration between the Pakistan stock market and the stock markets of India, China, Korea, Indonesia, Thailand, and Malaysia.

A study was performed by Nisa and Nishat (2011) to establish the nexus between the stock market and a large set of predictor variables including liquidity ratio, market to book value, capital structure, earning per share, dividend payout ratio, firm size, share turnover ratio, size of the stock market, GDP growth, interest rate, money, supply, financial depth, and inflation are all explanatory variables under this study. Excluding liquidity ratio, dividend payout ratio, and stock market stock size, all variables are significantly contributing to the stock market in Pakistan.

Shahbaz et al. (2015) attempted to discover the relevant factors of the stock market in case of Pakistan. By applying advanced statistical tools, they concluded that GDP, inflation, investment, trade openness, and financial development are the main drivers of the stock market. Eita (2012) explored the macroeconomic indicators of stock market prices in Namibia. The exploration was performed by implementing the VECM approach and inferring that the Namibian stock market is determined by inflation, GDP, interest rates, exchange rates, and money supply.

Saeed (2017) analyzed the association between government effectiveness, corruption, and political stability on stock market performance by utilizing the panel VEC mechanism for South Asian countries. The findings manifest a strong association between government effectiveness, control of corruption, political stability, and stock market performance. The study explored the influence of inflation, the exchange rate on stock market returns in Ghana. The ARDL approach was implemented to determine the relevant predictors. The study recommends the strong association between exchange rate, stock market, and inflation.

Asaad and Marane (2020) seek to explain how terrorism, corruption, oil prices, and political stability influence the Iraq stock market by using the ordinary least squares tool. The results show that the level of corruption, terrorism activities, political stability, and oil prices are significantly associated with the Iraq stock exchange. Papapetrou (2001) analyzed the influence of real stock prices, oil prices, interest rates, real economic activity and employment for Greece. The study applied VAR methodology and found that oil prices are significant in driving stock price movements.

Narayan and Narayan (2020) examined the impact of crude oil prices, nominal exchange rate, the growth rate in the nominal exchange rate, the growth rate of crude oil price, and the growth rate of the stock price on stock market prices. For this aim, they implemented an error correction model in the study. As a result, the study showed a strong association between crude oil prices, nominal exchange rate, and stock prices.

2.4. Rationale of the Study

The above-mentioned papers consider principal component analysis, independent component analysis, and sparse principal component analysis for the construction of the factor model.

However, there is also a small and growing body of literature investigating the classical approach (Autometrics) in the context of macroeconomic forecasting (Castle et al., 2013; Doornik and Hendry, 2015; Castle et al., 2020). We failed to discover any paper to date that has investigated the use of PLS theoretically in our context. However, the method has been applied empirically in various fields. Apart from this, some papers have utilized shrinkage methods like ridge regression, lasso, elastic net, adaptive lasso, and non-negative garrote, but none of the papers to date have used the updated forms of shrinkage methods in our context. Filling the gaps, this work implements some updated techniques of big data to increase the literature of macroeconomic forecasting theoretically as well as empirically. From the dimension reduction aspect, we build factor models with the intention of highlighting the importance of such models for macroeconomic prediction. In particular, while building factor models, we employ PCA and PLS. In addition, we also assess the last version of the classical approach (Autometrics) and the updated version of shrinkage methods, including E-SCAD and MCP, in terms of feature selection as well as forecasting. We evaluate the performance of these techniques using different data-generating processes.

To summarize the whole discussion, our prime contribution comes in the form of a comparison of Autometrics with updated shrinkage methods under the simulated scenarios having multicollinearity, heteroscedasticity, and autocorrelation along with their application to macroeconomic and financial datasets. Secondly, we compare the Autometrics and updated shrinkage methods with factor models under the same scenarios and real data to provide a conclusive solution to predictability. The study goal is to produce an improved method to help policymakers; the improved tool is not restricted to workers' remittances, stock market, or inflation (in our case) but is valid for any macroeconomic or financial data time series.

Chapter 3

Methodology

Model selection is one of the crucial steps of empirical research across all disciplines, where an earlier theory does not pre-define a proper specification. Economics is definitely one of them, as macroeconomic processes are generally high-dimensional, non-stationary, and very complex (Hendry and Krolzig, 2005). Typically, various solutions have been proposed to estimate the models. But picking a statistical model has become a very important and common part of empirical economic research.

Selection procedures such as information criteria, penalization tools, and stepwise regression are unavoidable. There can never be a consensus regarding which model is the best because there are a substantial number of criteria to evaluate the models' performance. Fortunately, over the last few decades, a new revolution has existed in model building, in the form of a general-to-specific approach, designated as gets. It is basically contained in a computer programme named PcGive. Computer automation of the Gets approach has provided fresh insight on how to choose a statistical model.

PcGive is a computer programme that automatically selects an econometric model. It is a completely new approach to formulating models and is particularly devised for handling economic data when the correct form of the equation under analysis is unknown. In PcGive, the automatic model selection job is performed by Autometrics. Hence, in the next section, we provide a detailed explanation of Autometrics.

3.1. Autometrics

The automated Gets procedure can almost be considered a "black box": a final model is chosen from the model that is constructed from an initial set of candidate variables. The initial model refers to the general unrestricted model (GUM). Mostly, a set of terminal candidate models is found. In such circumstances, information criteria are utilized as the tie-breaker. There is a possibility that we may choose the final GUM in the block-search procedure, which is the union of the terminal candidate models.

The aim of the automated gets procedure is to ensure that the GUM is well specified statistically, which is subjected to miss-specification testing. Hereafter, diagnostic tests guarantee that all underlying terminal candidate models clarify these tests as well. The simplification of GUM is done via path search. Such a type of search is needed to tackle the complex autocorrelation that is often present in macroeconomic data. A simplification is acceptable provided the expelled variables are insignificant and the new model is a good chopping of the GUM. This last condition is also called encompassing the GUM or backtesting, and it is based on the F-test of the removed variables in linear regression models.

In the application of Autometrics, reduction in p-value is the principal choice to be used for backtesting and individual coefficient significance. There are some tools to eschew estimating models (Doornik, 2009b). This method is very efficient, even though the costs of statistical inference cannot be circumvented and the costs of searching are substantially low. A pair of automatic model selection frameworks that fail to fit the model within the general to specific (gets) methodology are:

1. Stepwise regression: start with the empty model and add the most significant omitted variable in the model. The highly insignificant variable is removed from the model that is observed at any stage (hence in every iteration up to one predictor can be included, and one can be deleted) (Barrodale and Roberts, 1973). This method is repeated till we get all the variables in the model to be significant, and all omitted variables must be insignificant.

2. Backward elimination: all predictors are entered into the initial model, then predictors are thrown one at a time beginning with the least significant. The process is continued until all predictors have a p-value of p_a or small.

There are three main differences with automated versions of the above procedure: (i) lack of search, (ii) no backtesting; (iii) no miss-specification testing or diagnostic tracking. Figure 3.1 reveals how the Autometrics approach works for feature selection step by step. It starts with a general unrestricted model and arrives at a final model that contains all the significant variables, which refers to LDGP.

3.1.1. Methodology

Autometrics is comprised of five basic stages.

- In the first stage, the linear model known as the "so-called" General Unrestricted Model (GUM) is formed.
- In the second stage, the parameters are estimated along with the statistical significance of the GUM.
- In the third stage, the pre-search process is performed.
- The fourth stage produces the tree path-search.
- In the last stage, the final model is selected.

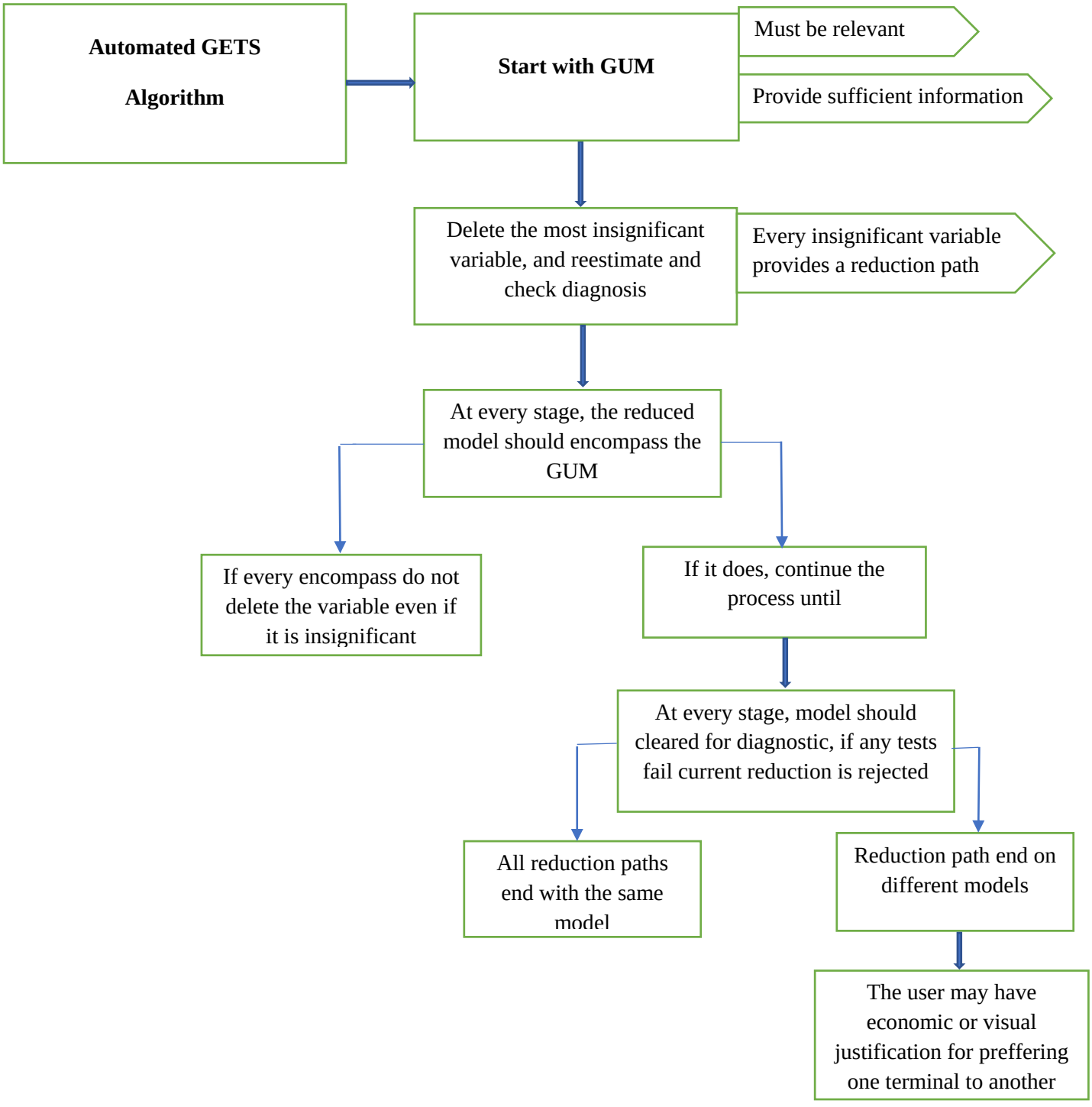


Figure 3.1: Major points of the Autometrics approach

Doornik (2009b) elaborated on the entire algorithm of Autometrics whereas the steps to run Autometrics are as follows. Start off by considering all the candidate variables in a linear model (GUM) and estimating them by the least-squares method, then verifying them through diagnostic tests. In case of insignificant coefficients, simpler models are estimated by utilizing a tree-path reduction search and validated by diagnostic tests. If some terminal models are detected, Autometrics undertakes its own union testing. Rejected models are deleted, and the union of those terminal models that survived induces a new GUM for another tree-path search iteration. This inspection procedure continues, and the terminal models are statistically assessed against their union. If two or more terminal models clear the encompassing tests, then the pre-chosen information criterion is a gateway to a final decision.

3.2. Shrinkage Methods

An alternative prominent approach to dealing with many features is the family of panelized regression methods, which comprises many techniques, but our study adopts the following updated forms: elastic smoothly clipped absolute deviation (ESCAD) and minimax concave penalty.

3.2.1. Elastic Smoothly Clipped Absolute Deviation (ESCAD)

The SCAD is non-convex and enjoys the oracle properties of sparsity, continuity, and unbiasedness. This technique selects useful covariates with their magnitudes asymptotically in an efficient way if the underlying true model is known, i.e., the oracle properties. The SCAD function covers all the limitations faced by existing methods like ridge and Lasso. The penalty function of SCAD is defined as:

$$p_k(|\tau|) = k\left\{I\ (\tau \leq k) + \frac{(\gamma k - \tau)}{(\gamma - 1)k} + I(\tau > k)\right\} \tag{3.1}$$

Where τ and γ are the two unknown parameters and in practice, the best pair of the dual parameters can be selected through generalized cross validation (GCV) and cross validation (Lu et al., 2014). The unknown tuning parameter *k* is determined by the generalized cross-validation approach, and the authors assumed the value of γ is 3.7, because this value works similarly to that selected by GCV approach (Fan and Li, 2001). As given above, the penalty function is continuous, and the resulting solution is given by:

$$p_k(|\tau|) = \begin{cases} k|\tau| & |\tau| < k \\ -(\tau^2 - 2\gamma k|\tau| + k^2)/2(\gamma - 1) & k < |\tau| \leq \gamma k \\ (\gamma + 1)k^2/2 & |\tau| > \gamma k \end{cases} \quad (3.2)$$

The tuning parameters can be induced by the data-driven technique. The limitation of SCAD is that it selects only one variable from a correlated set of predictors. Zeng and Xie (2014) extended the SCAD by augmenting L_2 penalty and called it elastic SCAD (E-SCAD). Mathematically, it can be written as:

$$pen_{\mathrm{k}}(|\tau|) = \sum_{d=1}^{D} p_{\mathrm{k}}(|\tau|) + \lambda_{2p} \sum_{d=1}^{m} {\alpha_d}^2 \quad (3.3)$$

In Equ. 3.3, the first term shows the SCAD penalty, and the second term is basically the ridge (L_2) penalty. Due to L_2 penalty, the E-SCAD achieves an additional property along with oracle properties, that is, the penalty function should spur highly correlated features to be in or out of the model simultaneously. Hence, the proposed form selects the whole group of correlated predictors rather than one variable.

3.2.2. Minimax Concave Penalty

Zhang (2010) proposed a minimax Concave Penalty (MCP), which increases the convexity of the penalized loss in sparse regions considerably given specific thresholds for feature selection as well as unbiasedness. The MCP is described as:

$$S_{MCP}(t;k) = \begin{cases} kt - \frac{t^2}{2\gamma} & if\ |t| \le \gamma k \\ \frac{1}{2}\gamma k^2 & if\ |t| > \gamma k \end{cases} \qquad (3.4)$$

Where both γ and k are the tuning parameters, and can be selected through cross validation or information criteria, namely, the Akaike information criterion (AIC) or Bayesian information criterion (BIC) (Breheny and Huang, 2011). The tuning parameter ($\gamma > 0$) diminishes the maximum concavity under the following restrictions like unbiasedness and selection of features:

$$\rho(t;k) = 0 \qquad \forall\ t \ge \gamma k \qquad \rho(0+;k) = k$$

$$\sum_{d=1}^{m} p_d\left(|\alpha_d|;k;\gamma\right)$$

The dual tuning parameters in concave penalty regression play a key role in terms of controlling the amount of regularization. Likewise, the concavity of the MCP penalty considerably evades the sparse convexity by diminishing the maximal concavity. In 2010, the author showed that a rise in regularization parameter values leads to more convexity and an almost unbiased penalties. The penalty function of MCP typically belongs to the quadratic spline function.

3.2.3. Adaptive Elastic net (AEnet)

The lasso estimator has been designed to improve the performance of the ridge estimator. It is certainly useful, particularly when most coefficients of the true model are zero. Albeit, ridge

regression performs better than lasso when a correlation between predictors is high (Zou and Hastie, 2005).

To overcome the shortcomings of lasso and ridge regression, the elastic net method was proposed by (Zou and Hastie, 2005) and used both lasso and ridge penalties simultaneously. The penalty function of the elastic net (EN) is given by:

$$\hat{\alpha}^{EN} = (1 + \frac{k_2}{n}) \text{ argmin } \sum_{c=1}^{n}(y_c - \alpha_o - \sum_{d=1}^{m} \alpha_d x_{cd}) + k_2 \sum_{d=1}^{m} \alpha_d^2 + k_1 \sum_{d=1}^{m} |\alpha_d| \quad (3.5)$$

Using a cross-validation approach, the tuning parameters k_1 and k_2 control the relative significance of L_1 norm and L_2 norm penalty. Both Lasso and Ridge regression, the special forms of the elastic net, have already been discussed in section 1.1. In this sense, the elastic net contains two features: shrinkage and variable selection. Besides, a detailed explanation of this aforementioned equation is given in section 1.1.

To estimate $\hat{\alpha}^{EN}$, (Zou and Hastie, 2005) proposed an algorithm called least angle regression (LAR). This is the fact that EN does not satisfy an oracle property like Adaptive Lasso, albeit it performs better than Adaptive Lasso (Algamal and Lee, 2015). Later on, the ideas of the Adaptive Lasso and the Elastic net regularization were combined to achieve further improvement known as Adaptive Elastic-net (AEnet) and is defined as:

$$\hat{\alpha}^{AEnet} = (1 + \frac{k_2}{n}) \text{ argmin } \sum_{c=1}^{n}(y_c - \alpha_o - \sum_{d=1}^{m} \alpha_d x_{cd}) + k_2 \sum_{d=1}^{m} \alpha^2{}_d + k_1 \sum_{d=1}^{m} \hat{\omega}_d |\alpha_d| \quad (3.6)$$

$\hat{\omega}_d$ *(d=1,2, …, m*) are adaptive data-driven weights. According to Zou and Zhang (2009), initially, we estimate the $\hat{\alpha}^{EN}$by using an EN method as given in Equ. (3.5) and then utilizing it while computing the weights as $\hat{\omega}_d = |\hat{\alpha}_d^{EN}|^{-\tau}$, here τ is constant and should be positive. Thus, AEnet, the modified form of elastic net, attains an oracle property.

3.3. Factor Models

The notion of factor models also called diffusion index entails the utility of properly extracted hidden common factors that have been distilled from a huge set of features as inputs in the identification of the parsimonious models. To be more specific, suppose X is a N × P dimensional matrix of data points and define N × k dimensional matrix of latent factors.

The forecasting tools, particularly, the factor models are delineated in depth by Stock and Watson (2006). In the below-detailed discussion of factor model methodology, we follow Stock and Watson (2002a):

$$X = F\,\varphi' + \varepsilon \tag{3.7}$$

Where ε represents the random error matrix, φ' is the P × k coefficients matrix and F is a factor matrix of N × k dimension.

We construct the following forecasting model based on the work of Bai and Ng (2006a), Kim and Swanson, (2014a) and Stock and Watson (2002a, b):

$$Y_{t+h} = F_t\gamma_F + e_{t+h} \tag{3.8}$$

Where Y_{t+h} is an outcome variable to be forecasted, h shows the forecast horizon, F_t is the vector of factors with a dimension, distilled from *F* in equation (3.7). The associated coefficient γ_F is a vector of unknown parameters and e_{t+h} is the random error.

3.3.1. Principal Component Regression (PCR)

The formulation of a factor-based model needed the following two steps. In the first step, we estimate k latent factors, let's say $\hat{F}$ by using P observable covariates. To gain convenient dimension reduction, k is supposed to be much smaller than P (i.e., k ≪ P). In the second step, we estimate $\hat{\gamma}_F$, by utilizing data at hand with Y_t and $\hat{F}_t$. Subsequently, an out-of-sample forecast is constructed.

Kim and Swanson (2014a) utilized the PCA approach to achieve estimates of the unobserved factors, known as principal components (PCs). The latent PCs are uncorrelated and obtained by using the data projection in the direction of maximal variance, and naturally, the PCs are ordered based on their variance contributions. The first PC reflects the direction of the maximal variance in the data, the second PC reflects the direction that explains the maximal variance in the rest of the orthogonal subspace and so on.

This approach is most frequently used in the literature of factor analysis because PCs are easily derived via the use of singular value decompositions (Stock and Watson, 2002a; Bai and Ng, 2002, 2006b). However, the performance of the factor model is more likely to be worse in the prediction if the incorporated factors are dominated by excluding factors (Boivin and Ng, 2006). Similarly, Tu and Lee (2019) stated that PCA imposes only the factor structure for X and does not consider the outcome variable. It indicates, no matter what the outcome to predict. By dint of neglecting the outcome variable at the time of factors, extraction induces an inefficient forecast of the outcome variable. The solution to this problem is given in the next section.

3.3.2. The Partial Least Squares (PLS) Method

This study looks at another method known as partial least squares (PLS) regression developed by Wold (1982). This method is appropriate in a data-rich environment and may be considered as an alternative to PCA-based factor models. Unlike the PCA method, the PLS identifies new factors in a supervised way that is, it makes use of the response variable to identify new factors that not only approximate the old factors well but are also related to the response variable. Roughly speaking, the PLS approach attempts to find the directions of maximum variance that help explain both the response variable and explanatory variables. The PLS for an outcome variable is motivated by a statistical model as follows:

$$Y_t = x_t\gamma_P + e_t \tag{3.9}$$

Where $x_t = [x_{1,t}, x_{2,t}, \dots, x_{n,t}]'$ is n × 1 vector of covariates at time *t = 1, …, T*, γ_P is n × 1 vector of associated coefficients, and e_t is the disturbance term. Kim and Ko (2020) argued that PLS type models are useful especially when there are a large number of covariates. Instead of using a model given in (3.7), one may adopt another data dimension reduction approach through the following linear regression with *Z* × 1 vector of components $s_t = [s_{1,t}, s_{2,t}, \dots, s_{Z,t}]$ as follows;

$$Y_t = x_t w\tau + e_t \tag{3.10}$$

$$Y_t = s_t\tau + e_t$$

We define s_t,

$$s_t = w'x_t$$

Where $w = [x_1, x_2, \dots, x_Z]$ is the *n* × Z matrix of each column, $w_z = [w_{1,z}, w_{2,z}, \dots, w_{n,z}]'$, z = 1, 2, …, Z, denote the vector of weights on covariates for z factors or components and τ is the Z × 1 vector of PLS coefficients. We may use the following equation for predicting the k steps ahead model that is $\hat{y}_{t+k}$, *k* = 1, 2, …, *m*.

$$\hat{y}_{t+k} = \hat{\gamma}_k' x_t \tag{3.11}$$

3.4. Selection of Tuning Parameter(s)

A cross-validation strategy is frequently used to determine the tuning parameter in order to achieve the best prediction solution. It requires randomly dividing the input data into two halves: a training data set and a testing data set (or hold-out set). The training data set is used to fit the model, and the fitted model is used to predict answers for the validation data set. The validation set error rate,

which is often determined using MSE in the context of a numerical answer, is used to estimate the test error rate. Using a k-fold CV, the K-fold cross-validation method randomly divides data collection into k groups, or folds, of roughly similar size; often, k = 10 or 5.

To achieve an accurate prediction, the cross-validation process is frequently used for the selection of tuning parameters, also called hyperparameters. In general, it requires the data to be partitioned into two parts; training dataset and testing dataset (validation set). The former part is utilized for model fitting and then this estimating model is utilized for prediction for the validation set. The test error is obtained by using validation set error rate, which is usually computed in the form of a mean square error (MSE). The k-fold cross validation (CV) approach involves the random splitting of data into k folds or categories of the same size, utilizing a k-fold CV. In this study, 10-fold cross-validation is executed to determine the optimal value of the tuning parameter(s). The remaining data is used for model fitting, with the first fold acting as a validation set, where the MSE_i is computed on the held out fold. This process is repeated k times, with each validation set involving a distinctive set of legs (observations). In this way, the test error is estimated as MSE_1, MSE_2, ..., MSE_k. The k-fold CV estimate is achieved by averaging the values of MSE_i.

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i \tag{3.12}$$

3.5. Simulation Study

In statistics and econometrics, it is imperative to investigate the performance of statistical models theoretically and empirically. Empirically, nobody knows the true data generating process, and the researchers often fail to determine whether the underlying method retains the correct variables or not. On the other hand, the true DGP is known in the Monte Carlo simulation experiments, which enables us to compare different econometric techniques and arrive at correct conclusion. The one

that is close to the true DGP is the winner (preferenced comparatively). This study seeks to describe both aspects of advanced statistical and machine learning techniques. Our simulation experiment involves three main scenarios, namely simulations on a data generating process (DGP) with (i) multicollinearity, (ii) heteroscedasticity, (iii) autocorrelation. In each case, we vary the DGP characteristics such as the correlation structure among predictors, the level of variance of the error term, and the level of correlation between the current and lagged value of the error term. Simulation as a method of analysis has the main disadvantage of being specific to the setup.

3.5.1. Data Generating Process

The question we try to answer, which procedure is best in feature selection as well as forecasting using big data in cases of non-orthogonal structure. First, we write the model, which is general data generating process and later try to match it with the specific assumptions.

$$Y = X_i\beta + \mu \tag{3.13}$$

The set of predictors X_1, X_2,…, X_P are generated from multivariate normal distribution as $X_i \sim N(0, \sum)$. The same data generating process (DGP) was used by (Wahid et al., 2017; Smeekes and Wijler, 2018) as mentioned in (3.13) for artificial data generation.

In this DGP, the total number of predictors (candidate variables) is '*P*'. We have to divide the predictor set into two parts, i.e., relevant and irrelevant variables. The relevant variables are merely used to generate the DGP of *Y* and the coefficients of irrelevant variables are set to zero.

In this study, we are interested using different scenarios such as multicollinearity, heteroscedasticity, and autocorrelation, and for each scenario, a separate procedure for DGP is illustrated below.

3.5.1.1. SCENARIO-I
Candidate set of features are correlated

In this scenario, we consider the DGP, where the set of features are correlated (non-orthogonal) with each other. In real world phenomena, inclusion of feature in the model is based on two criteria; (i) correlated with response variable and, (ii) correlated with the included covariate in the model. Considering the second condition, the correlation should be weak, but in fact, it is not always the case. Particularly in economics and finance, the set of predictors is often highly associated with each other in a rich data environment. The presence of moderate or high multicollinearity adversely influences the inference and prediction of the estimated coefficients (Ali et al., 2021). Thus, in a simulation exercise, it is important to concentrate on the different levels of correlation among covariates and discover the performance of the aforementioned methods to hold the true DGP.

The procedure of DGP is same as discussed in (3.13), except, that here we generate the pairwise correlation between the predictors i.e., x_m and x_n as $cov(x_m, x_n) = \sum^{|m-n|}$. The population covariance matrix is produced in the following way:

$$\sum_P = \begin{bmatrix} 1 & . & . & . & \sum^{|n-m|} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \sum^{|m-n|} & . & . & . & 1 \end{bmatrix}$$

It is a fact that the variance-covariance matrix contains variance and covariance together. However, by altering the parameter $\sum_P$ we obtain different correlation structures. In our work, we assume values for $\sum_P \in \{0.25, 0.5, 0.9\}$ as followed by Xiao and Xu (2015).

3.5.1.2. SCENARIO-II
Error term is Heteroscedastic

Like multicollinearity and autocorrelation, heteroscedasticity is also one of the main problems of OLS estimators. In presence of heteroscedasticity, the OLS estimators remain unbiased and

consistent but their efficiency (standard error) is negatively influenced. In other words, we can say that the OLS estimators are no longer the best linear unbiased estimators (BLUE) but remain only linear unbiased estimators (LUE). In such circumstances, the OLS produces invalid statistical tests like 't' and 'F' i.e., we cannot achieve satisfactory results.

In this scenario, we set the DGP in such a way that error terms are generated as heteroscedastic. More specifically, here we rely on the examination of heteroscedasticity i.e., that the variance of the error term is generated non constantly and alters across data points by σ_k. In the real world, it is not feasible to find the orthogonal structure of variables, therefore, in the same DGP, the set of regressors is introduced as correlated (moderate case of multicollinearity i.e., $\sum_P = 0.5$).

$$E(\mu_t^2) = \sigma_k^2 \tag{3.15}$$

In equ. (3.15), we split the variance $\boldsymbol{\sigma^2}$ of the error term into two components i.e., $\boldsymbol{\sigma_1^2}$ and $\boldsymbol{\sigma_2^2}$. Let we have 'n' observations, and divide them into two parts as n_1 and n_2. We set the variance of n_1 (first part of data) as $\boldsymbol{\sigma_1^2}$ and the variance of n_2 (second part of data) as $\boldsymbol{\sigma_2^2}$, followed the same procedure by Khan (2022). Our simulation experiments assume three different cases of heteroscedasticity as following:

$\pi_i = (\boldsymbol{\sigma_1/\sigma_2})$, where i = 1, 2, 3 as $\pi_i \in \{0.1/0.3, 0.2/0.6, 0.3/0.9\}$.

3.5.1.3. SCENARIO-III
Error Term is Autocorrelated

Under various circumstances, the researchers fail to fit the model that is specified correctly. In other words, it can be said that miss-specification is a very common issue, which researchers often face, particularly, in economics. The miss-specified model produces an issue of autocorrelation, which is always problematic. Moreover, measurement error is also more likely to be observed in the variables, which can cause the problem of autocorrelation. Using the same model (3.13), we

generate the correlation between current and residual lag (autocorrelation) and symbolized by ρ. Mathematically, the autocorrelation is generated as:

$$\mu_t = \rho\mu_{t-1} + \varepsilon_t \tag{3.16}$$

Where

$$\varepsilon_t \sim N(0, 1)$$

Our experiments assume low, moderate, and high cases of autocorrelation such as $\rho \in \{0.25, 0.5, 0.9\}$. All three cases of autocorrelation are examined for different sample sizes and different sets of candidate variables (relevant and irrelevant). As disclosed in the recent last scenario that orthogonal set of regressors is not feasible for the same response variable, however in addition to autocorrelated errors, we introduce the moderate multicollinearity among regressors in the DGP (because in the real world, the set of predictor variables are often correlated each other, particularly in the field of economics and finance).

3.5.2. Measures of Methods Performance

There are a few ways to evaluate the models' performance in terms of variable selection, in which we are adopting the potency and gauge. Gauge is delineated as the empirical null retention frequency that how often irrelevant covariates are retained. The comparison of Autometrics with penalization methods is evaluated in the form of correct zero identification interpreted as potency and incorrect zero identification referred to as Gauge (Doornik and Hendry, 2015).

Mathematically, the gauge can be expressed as:

$$\hat{p}_{irr}/p_{irr}$$

$$E\left(\frac{\hat{p}_{irr}}{p_{irr}}\right) \rightarrow \alpha$$

The gauge indicates the irrelevance part which corresponds to nominal significance level (α), where p_{irr} shows a set of irrelevant covariates in the initial model and $\hat{p}_{irr}$ shows the set of estimated irrelevant covariates (Pretis et al., 2018).

Potency is defined as;

$$\hat{p}_{rel}/p_{rel}$$

$$E\left(\frac{\hat{p}_{rel}}{p_{rel}}\right) \rightarrow 1$$

This indicates that the relevant part p_{rel} shows the set of relevant covariates in the initial model and $\hat{p}_{rel}$ point to the set of estimated relevant covariates, so the expected potency tending towards the value 1 is evidence of a good model (Pretis et al., 2018). Furthermore, we repeat each simulation experiment 1000 times and the expected potency and gauge evaluate the best method relatively. For analysis, we have relied on several packages like gets, glmnet, ncvreg, pls, caret. forecast and Metrics under the R programming language.

To compare the predictive abilities of all procedures, we split the data set in such a way that 80 percent of the data is utilized for models' training and the remaining data are utilized for models' assessment. We repeat the process H = 1000 times. The average of root mean square error (RMSE) and mean absolute error (MAE) is calculated over 'H' to evaluate the predictive performance. Through these two criteria, we can achieve the prediction accuracy of all methods. The smaller values of MAE and RMSE indicate a better forecast comparatively. Their mathematical expressions can be illustrated as

$$\text{MAE} = mean\left(\left|Y_t - \hat{Y}_t\right|\right) \tag{3.17}$$

$$\text{RMSE} = \sqrt{mean\left(Y_t - \hat{Y}_t\right)^2} \tag{3.18}$$

In equations (3.17) and (3.18), Y_t and $\hat{Y}_t$ indicate the actual and forecast values respectively.

At the end, to summarize the entire chapter, three different aspects are discussed: methods, simulation setup and measures of method evaluation. The next chapter consists of two main subsections: variable selection and out-of-sample forecasting. In the first subsection, we evaluate and compare the performance of machine learning/shrinkage methods with Autometrics using Huge Big data. For that purpose, first we perform the simulation experiments under different conditions, namely multicollinearity, heteroscedasticity and autocorrelation with varying sample sizes and sets of candidate variables (relevant and irrelevant variables) to evaluate the underlying methods. After completing the simulation exercises, we check their performance using real dataset as well. In the second subsection, we seek to assess the predictive power of the proposed factor model based on PLS against various existing techniques. To achieve this, we have to perform simulation studies under the aforementioned scenarios, and then carry out a real data analysis for evaluating the models.

Chapter 4

Results and Discussion

In order to achieve the first goal of the study, we divided this chapter into two major subsections. The first subsection provides a comprehensive analysis of the automatic model selection tools, including E-SCAD, MCP, AEnet and Autometrics in different DGPs using huge big data **(P<N)**. We also provide their systematic comparison in terms of feature selection, the mainstream tools utilized in both Monte Carlo simulations as well as economics and finance. Similarly, the second subsection yields an extensive inspection of the earlier mentioned tools in different DGPs using fat big data **(P>N)**. In practice (real world phenomena), we do not know which subset of predictors is important for "y" (response variable). Thus, it is hard to do this kind of practice directly with real data. The simulation experiments let us compare different statistical tools in different DGPs and find out how well they work in certain situations.

Statistical learning has two fundamental goals: ensuring high prediction accuracy and discovering relevant predictive features. Selection of features selection is particularly important when the true underlying model has a sparse representation. Identifying significant features will raise the prediction power of the estimated model (Zou, 2006).

4.1. Comparison of Feature Selection Procedures using Huge Big Data

There is a list of tools that exist in the literature for feature selection, but the current study only focuses on the updated versions of these tools. In general, we compare their performance under different scenarios where the number of observations is more than the number of covariates.

4.1.1. Design of Experiments and Results

For simulation experiments, three kinds of scenarios are considered by allowing the multicollinearity, heteroscedasticity, and autocorrelation conditions with varying sample sizes and a varied number of covariates. To be more specific, the random finite samples of sizes 80, 160, and 320 are drawn from the Gaussian distribution with 1,000 times replications. Moreover, we assume two sets of candidate variables with varying numbers of relevant (p) and irrelevant predictors (q) respectively, as presented in Fig. 4.1.

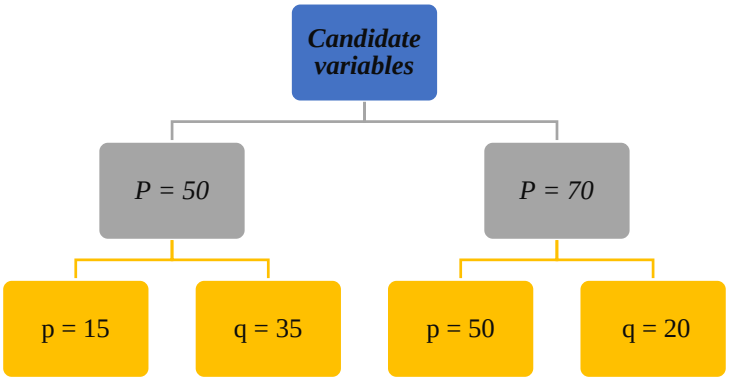


Figure 4.1: Distribution of candidate variables into relevant (p) and irrelevant (q)

We set the true parameters for P=50 and P=70, including intercept as,

$$\beta = (3, \underbrace{1, \dots, 1}_{15}, \underbrace{0, \dots, 0}_{35})$$

$$\beta = (3, \underbrace{1, \dots, 1}_{20}, \underbrace{0, \dots, 0}_{50})$$

Here, we check the performance of variable selection procedures under two different sets of regressors. In the dual sets, we set the intercept to be 3, and the candidate set of regressors is halved into relevant and irrelevant variables. The first set contains 15 relevant variables, to which we assign the value 1 (to each variable) and 35 irrelevant variables, having no effect on DGP (we

assign zero to the coefficient of each variable), whereas the second set contains 20 relevant and 50 irrelevant variables. A detailed explanation regarding DGP is provided in the preceding chapter. In order to evaluate the performance of all tools, two criteria will be used, i.e., potency and gauge. The Monte Carlo simulation results are discussed in Tables 4.1-4.3.

4.1.2. SCENARIO-I
The candidate set of features is correlated

This scenario considers the DGP with a set of correlated covariates. In real-world phenomena, very often, collinearity among the set of regressors is observed, whether it is weak, moderate, or strong. In such circumstances, it is crucial to estimate the desired effect of the unknown parameter through the OLS approach (Ali et al., 2021). Thus, in simulation experiments, it is of great importance to focus on the various levels of correlation among the set of covariates and discover the performance of variable selection methods to hold the true DGP. Furthermore, the details regarding the DGP are given in the previous chapter.

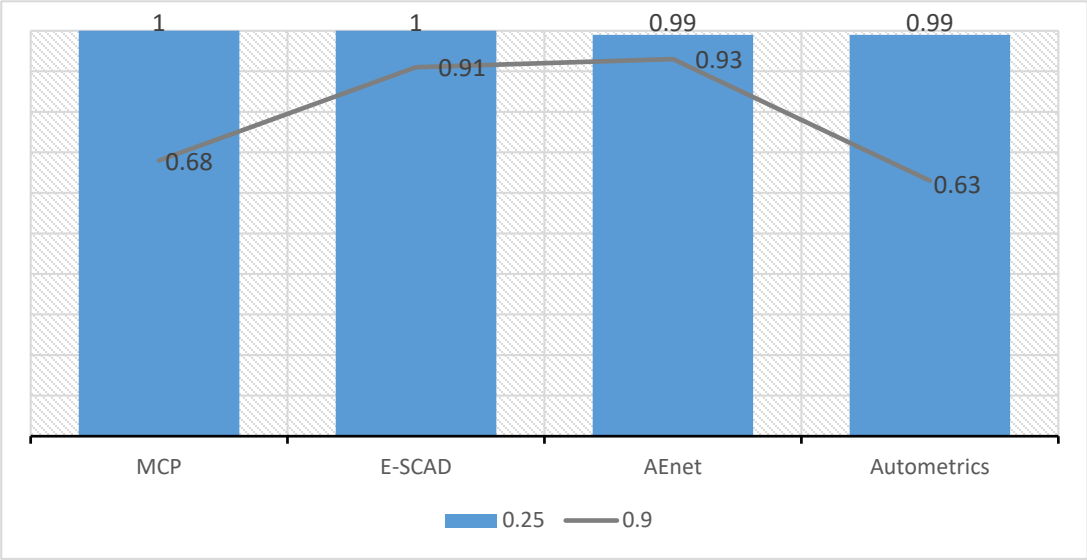
Table 4.1 depicts the findings of simulation in the cases of low, moderate, and high multicollinearity for different combinations of observations (n) and covariates. The performance of all methods improves with increasing sample sizes. In case of low multicollinearity, the potency associated with all methods is one under most simulated scenarios, clearly revealing that they retain all the relevant variables under low multicollinearity, as depicted by Fig. 4.2(a). The MCP and Autometrics keep 4% of irrelevant variables, AEnet keeps 1% of irrelevant variables, and E-SCAD keeps 12% of irrelevant variables. It means that E-SCAD substantially over-specifies the model by retaining a huge set of irrelevant variables. As the number of variables gets longer (p = 70), we do not see any big changes in the models' results.

The moderate level of multicollinearity does not adversely influence the potency. In terms of gauge, it tends to improve the performance of AEnet and Autometrics in such a way that they hold fewer irrelevant variables, but adversely affects the MCP performance, particularly in a small sample. The gauge associated with E-SCAD has been considerably improved. When we take the number of candidate variables, i.e., p = 70, the gauge of MCP and Autometrics deteriorate under the small size.

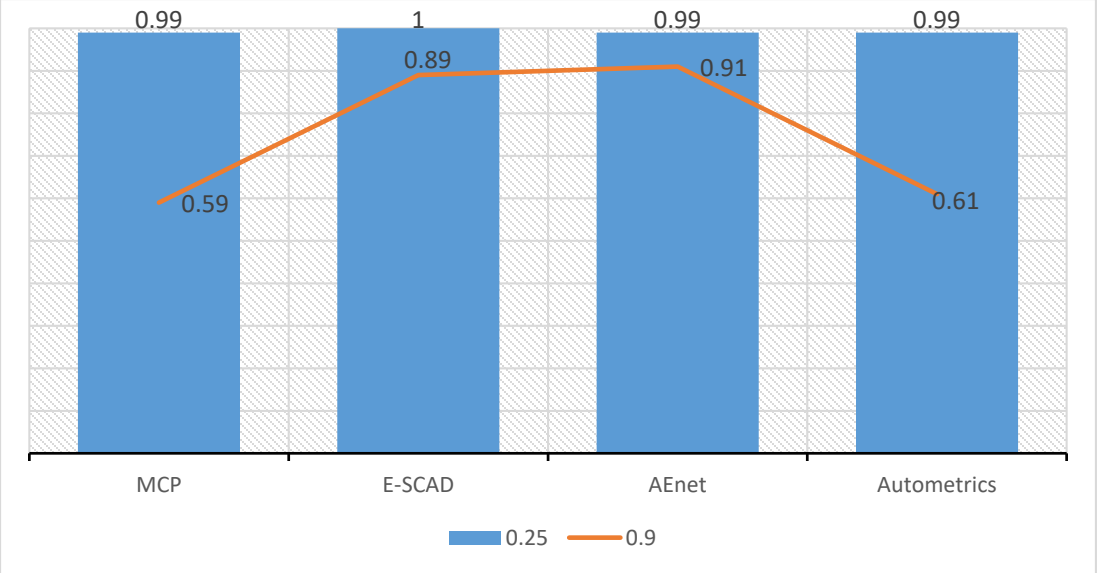
As shown in Figure 4.2(a,b), the high collinearity among the set of covariates substantially distorts the performance of all methods. More specifically, the gauges associated with MCP, E-SCAD and Autometrics have significantly deteriorated, as portrayed in Fig. 1 (see Appendix B). But as we increase the sample size, the E-SCAD and Autometrics become stable in terms of potency and gauge. In comparison, the AEnet holds more than 93 percent of the correct variables with a perfect gauge (zero percent).

Table 4.1. Variable Selection under Multicollinearity from Monte Carlo Simulation

Models	$\sum$ ***= 0.25, P = 50***		$\sum$ ***= 0.25, P = 70***	
n = 80/160/320	**Potency**	**Gauge**	**Potency**	**Gauge**
MCP	1/1/1	0.04/0.02/0.02	0.99/1/1	0.05/0.02/0.01
E-SCAD	1/1/1	0.12/0.10/0.10	1/1/1	0.11/0.10/0.09
AEnet	0.99/1/1	0.01/0/0	0.99/1/1	0.02/0/0
Autometrics	0.99/1/1	0.04/0.01/0.01	0.99/1/1	0.04/0.01/0.01
n = 80/160/320	$\sum$ ***= 0.50, P = 50***		$\sum$ ***= 0.50, P = 70***	
MCP	0.99/1/1	0.06/0.02/0.01	0.99/1/1	0.09/0.01/0.01
E-SCAD	1/1/1	0.10/0.07/0.06	0.99/1/1	0.09/0.06/0.06
AEnet	0.99/1/1	0/0/0	0.99/1/1	0/0/0
Autometrics	0.99/1/1	0.02/0.01/0.01	0.98/1/1	0.06/0.01/0.01
n = 80/160/320	$\sum$ ***= 0.90, P = 50***		$\sum$ ***= 0.90, P = 70***	
MCP	0.68/0.94/0.99	0.19/0.22/0.09	0.59/0.92/0.99	0.16/0.23/0.09
E-SCAD	0.91/0.98/0.99	0.13/0.09/0.03	0.89/0.98/0.99	0.12/0.09/0.03
AEnet	0.93/0.98/0.99	0/0/0	0.91/0.98/0.99	0/0/0
Autometrics	0.63/0.89/0.99	0.06/0.02/0.02	0.61/0.87/0.99	0.17/0.03/0.01



(a)



(b)

Figure 4.2: Potency under low and high cases of multicollinearity, when n = 80 and P = 50(a) and P = 70(b).

4.1.3. SCENARIO-II
Error Variance is Heteroscedastic

Heteroscedasticity seriously affects the OLS estimation process. In presence of heteroscedasticity, the OLS estimators are consistent and unbiased but suffer from high standard errors. In other words, it can be inferred that the OLS estimators are solely LUE (linear unbiased estimator) and do not remain BLUE. In such situations, the t and F tests are unreliable and do not provide satisfactory results. The DGP is already elaborated in the preceding chapter.

Table 4.2 presents the simulation results by varying heteroscedasticity along with sample size and a candidate set of variables (both relevant and irrelevant).

In case of heteroscedastic errors, the potency of all included methods is one in almost all scenarios, which certainly manifests that they hold all the active covariates. In terms of gauge, the MCP and E-SCAD keep more inactive variables and thereby over-specify the model. As we increase the sample size, the gauge of E-SCAD dramatically decreases. Similarly, AEnet and Autometrics avoid irrelevant variables and very precisely identify the true model.

Table 4.2. Variable selection under Heteroscedasticity from Monte Carlo Simulation

Models	π_1 **= 0.1/0.3,** ***P = 50***		π_1 **= 0.1/0.3,** ***P = 70***	
n=80/160/320	**Potency**	**Gauge**	**Potency**	**Gauge**
MCP	1/1/1	0.08/0.02/0.01	1/1/1	0.01/0.01/0.01
E-SCAD	1/1/1	0.10/0.11/0.11	1/1/1	0.09/0.10/0.10
AEnet	1/1/1	0/0/0	1/1/1	0/0/0
Autometrics	1/1/1	0.01/0.01/0.01	1/1/1	0.04/0.01/0.01
n=80/160/320	π_2 **= 0.2/0.6,** ***P = 50***		π_2 **= 0.2/0.6,** ***P = 70***	
MCP	1/1/1	0.02/0.01/0.02	1/1/1	0.01/0.01/0.01
E-SCAD	1/1/1	0.10/0.10/0.12	1/1/1	0.09/0.10/0.10
AEnet	1/1/1	0/0/0	1/1/1	0/0/0
Autometrics	1/1/1	0.01/0.01/0.01	1/1/1	0.04/0.01/0.01
n=80/160/320	π_3 **= 0.3/0.9,** ***P = 50***		π_3 **= 0.3/0.9,** ***P = 70***	
MCP	1/1/1	0.02/0.01/0.02	1/1/1	0.01/0.01/0.01
E-SCAD	1/1/1	0.10/0.10/0.10	1/1/1	0.09/0.10/0.10
AEnet	1/1/1	0/0/0	1/1/1	0/0/0
Autometrics	1/1/1	0.01/0.01/0.01	0.99/1/1	0.04/0.01/0.01

4.1.4. SCENARIO-III

The error term is Autocorrelated (moving average)

In many cases, the researchers fail to fit the model which is accurately specified. In general, we can say that miss-specification is a very frequent problem that researchers often face, especially in the field of economics. The miss-specified model leads to the problem of autocorrelation, which is always problematic. Probably, the existence of measurement errors also causes the problem of autocorrelation.

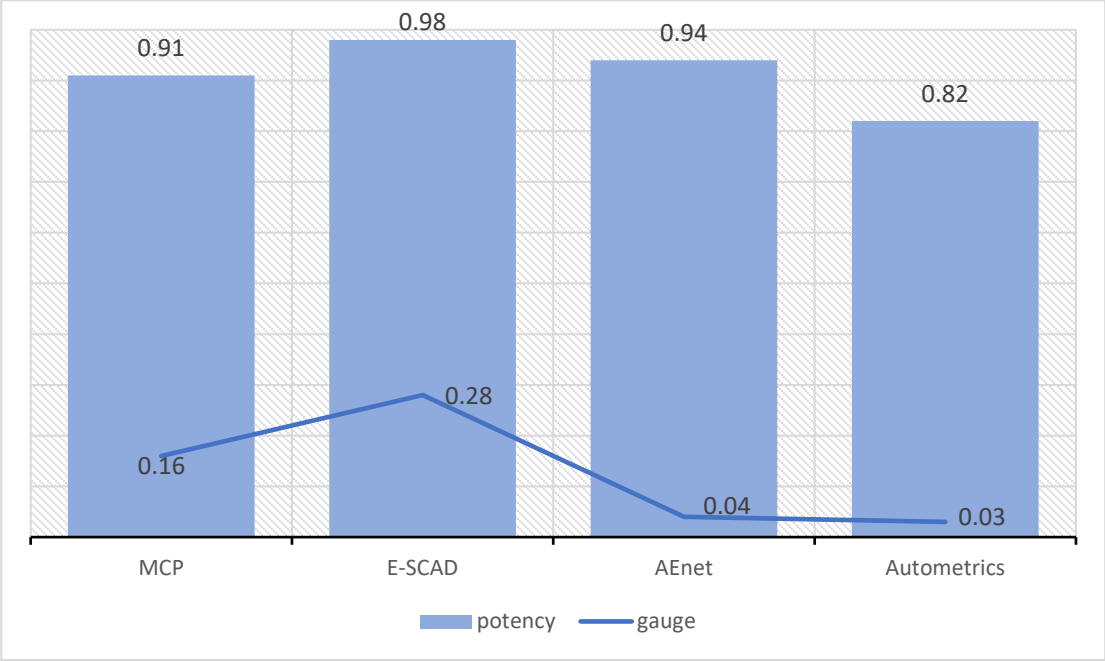
Table 4.3 portrays the simulation's output by varying Autocorrelation, sample size, and several covariates (both active and inactive). Low (0.25), moderate (0.5) and high (0.9) autocorrelation levels are considered here. In case of low and moderate Autocorrelation, the methods have often found all the right variables, but E-SCAD and MCP retain a huge set of irrelevant variables and thereby over-specify the model. In contrast, the AEnet and Autometrics provide the best results under almost all combinations of n and p. In other words, AEnet and Autometrics avoid the irrelevant variables and correctly specify the true model. By increasing the length of covariates, the E-SCAD gauge is slightly improved but adversely affects the gauge of Autometrics and AEnet. In the same way, it has a negative impact on the MCP gauge, particularly if there is a low autocorrelation case.

In the case of high Autocorrelation, the potency of E-SCAD is close to one and shows satisfactory performance, but the potency of the competitive methods is far away from one, which demonstrates that they miss the important variables. The same method i.e., E-SCAD collapsed under gauge. As shown in Fig 4.3(a, b), Autometrics and AEnet performed better in gauge and frequently held less than 5% of inactive variables. Expanding the covariates' window adversely

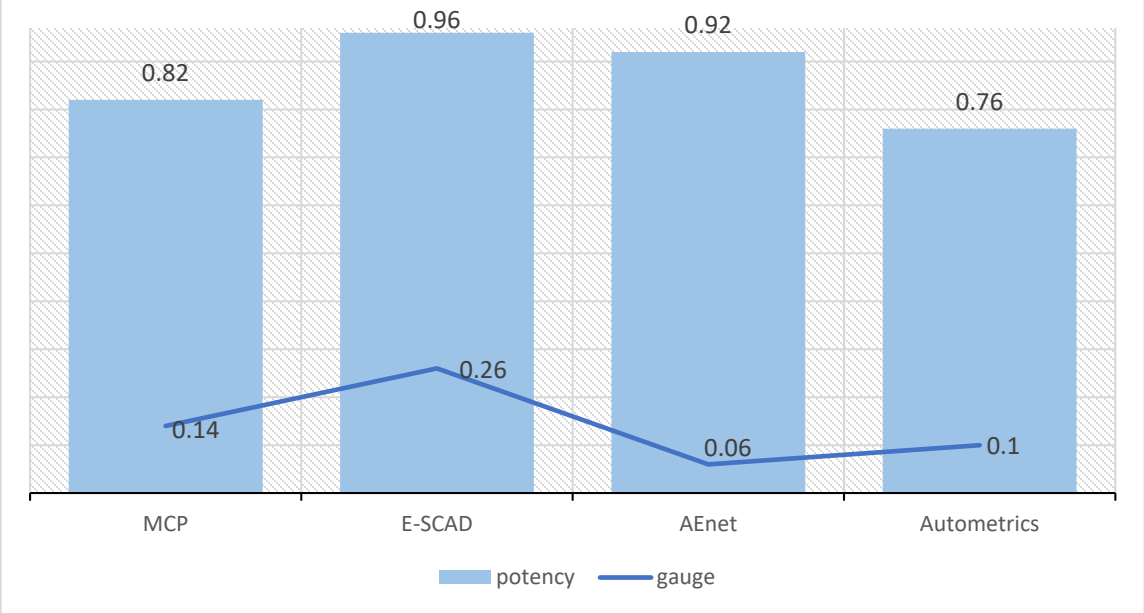
affects AEnet and Autometrics performance in terms of gauge, but has a positive influence on E-SCAD and MCP.

Table 4.3. Variable selection under Autocorrelation from Monte Carlo Simulation

Models	***ρ* = 0.25, *P* = 50**		***ρ* = 0.25, *P* = 70**	
n=80/160/320	**Potency**	**Gauge**	**Potency**	**Gauge**
MCP	1/1/1	0.04/0.02/0.02	1/1/1	0.04/0.02/0.02
E-SCAD	1/1/1	0.13/0.10/0.10	1/1/1	0.12/0.09/0.09
AEnet	0.99/1/1	0.01/0/0	0.99/1/1	0.02/0/0
Autometrics	0.99/1/1	0.01/0.01/0.01	0.99/1/1	0.05/0.01/0
n=80/160/320	***ρ* = 0.50, *P* = 50**		***ρ* = 0.50, *P* = 70**	
MCP	0.99/1/1	0.06/0.02/0.02	0.99/1/1	0.08/0.02/0.01
E-SCAD	1/1/1	0.15/0.10/0.10	0.99/1/1	0.14/0.09/0.09
AEnet	0.99/1/1	0.02/0/0	0.99/1/1	0.03/0/0
Autometrics	0.99/1/1	0.01/0.01/0.01	0.99/1/1	0.05/0.01/0.01
n =80/160/320	***ρ* = 0.90, *P* = 50**		***ρ* = 0.90, *P* = 70**	
MCP	0.91/0.99/1	0.16/0.12/0.05	0.82/0.99/1	0.14/0.11/0.05
E-SCAD	0.98/0.99/1	0.28/0.23/0.15	0.96/0.99/1	0.26/0.22/0.13
AEnet	0.94/0.99/0.99	0.04/0.01/0	0.92/0.99/0.99	0.06/0.01/0
Autometrics	0.82/0.98/0.99	0.03/0.01/0.01	0.76/0.97/0.99	0.10/0.01/0.01



(a)



(b)

Figure 4.3: Potency and Gauge across when ρ = 0.90, n = 80 and P = 50(a) and P = 70(b).

4.2. Comparison of Feature Selection Procedures using Fat Big Data

The development of econometric tools presents various challenges to modern data. The first challenge comes from the large data environment. Fat big data, where the length of predictors (p) greatly exceeds the number of observations (n) (Ye et al., 2021). For example, the volume of the dataset observed in economic research and application grows very swiftly and is more likely to affect economic policies as well as other economic activities (Eisenstein and Lodish, 2002). Hence, economic data indicates a useful asset with an under-utilized opportunity for the formulation of economic policy and its importance for the economic and social state of the nation. There is a wide range of sources that have economic and financial data with a huge set of features, including retail, real sector (output), prices, online trade, insurance, advertising, risk management, portfolio optimization, effect of education on earnings, labor market dynamics, money, exchange rates, interest rates, fiscal sector, and stock market dynamics (Syed and Lee, 2021, Belloni et al., 2013; Fan et al., 2014). So, the accurate analyses of economic data in the data-rich environment (many predictors) has become an emerging problem in the recent era of advanced econometrics.

This section analyzes the performance of variable selection methodologies, previously described in section 4.1, under different simulation experiments utilizing the Fat Big data. We also provide a systematic comparison of all these tools in terms of feature selection.

4.2.1. Design of Experiments and Results

For the simulation experiments, three types of scenarios are examined by allowing the Multicollinearity, Heteroscedasticity, and Autocorrelation problems to be altered by altering the number of predictors and sample size. We split the two candidate sets of predictor variables into 50 and 70, and further divided them into relevant (p) and irrelevant (q) predictor variables, as

depicted in Fig. 4.4. Aside from this, the random samples of sizes 40, 80, and 100 are drawn 1,000 times from a Gaussian distribution.

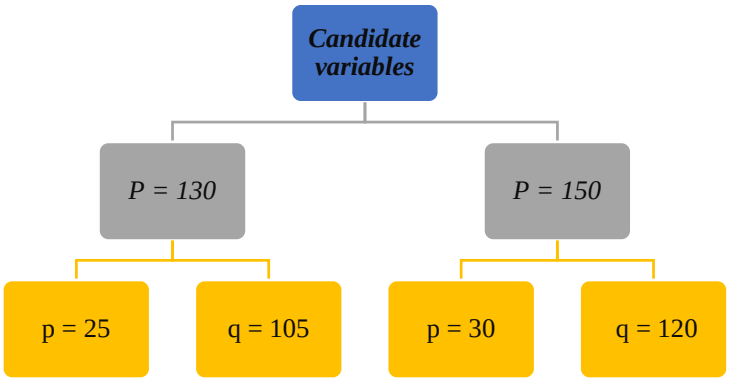


Figure 4.4: Distribution of candidate variables into relevant and irrelevant variables.

We set the true parameters, including intercept in the following way:

$$\beta = (3, \underbrace{1, \ldots, 1}_{25}, \underbrace{0, \ldots, 0}_{105})$$

$$\beta = (3, \underbrace{1, \ldots, 1}_{30}, \underbrace{0, \ldots, 0}_{120})$$

In the above, the intercept is set to be 3; the first set assumes 25 relevant and 105 irrelevant variables, and the second set assumes 30 relevant and 120 irrelevant variables while generating the DGP. Further expalanation regarding the DGP is provided in the preceding chapter. The results of the Monte Carlo simulation experiments are provided in Tables 4.4-4.6.

4.2.2. SCENARIO-I

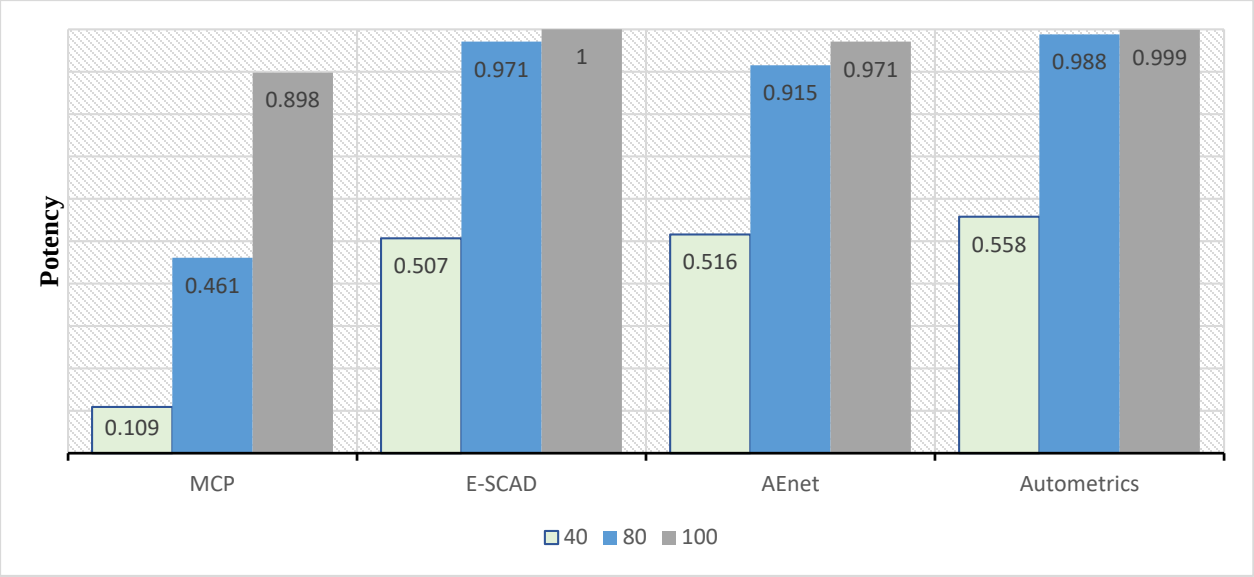
Table 4.4 depicts the findings of simulation in the case of low, moderate, and high multicollinearity for different combinations of observations (n) and covariates. The potency of all methods is improving with increasing sample size, as depicted in Fig. 4.4(a, b).

In case of low multicollinearity, the potency associated with Autometrics is comparatively high under n = 40, but with expanding the sample size, the potency of E-SCAD switches to one. As we take into consideration more covariates (P =150), the performance of MCP and E-SCAD improved, whereas the Autometrics and AEnet were adversely influenced. In terms of gauge, the Autometrics performance is best amongst the competitors. Furthermore, in Fig 4.5(a, b), it is noted that the MCP potency is very low at n = 40, but as the sample size increases, its potency tends to rise rapidly against the rival methods. The moderate level of multicollinearity negatively affects the potency of Autometrics and AEnet, and enhances the potency of MCP and E-SCAD. With expanding the covariate window to 150, all methods gain improvement in potency.

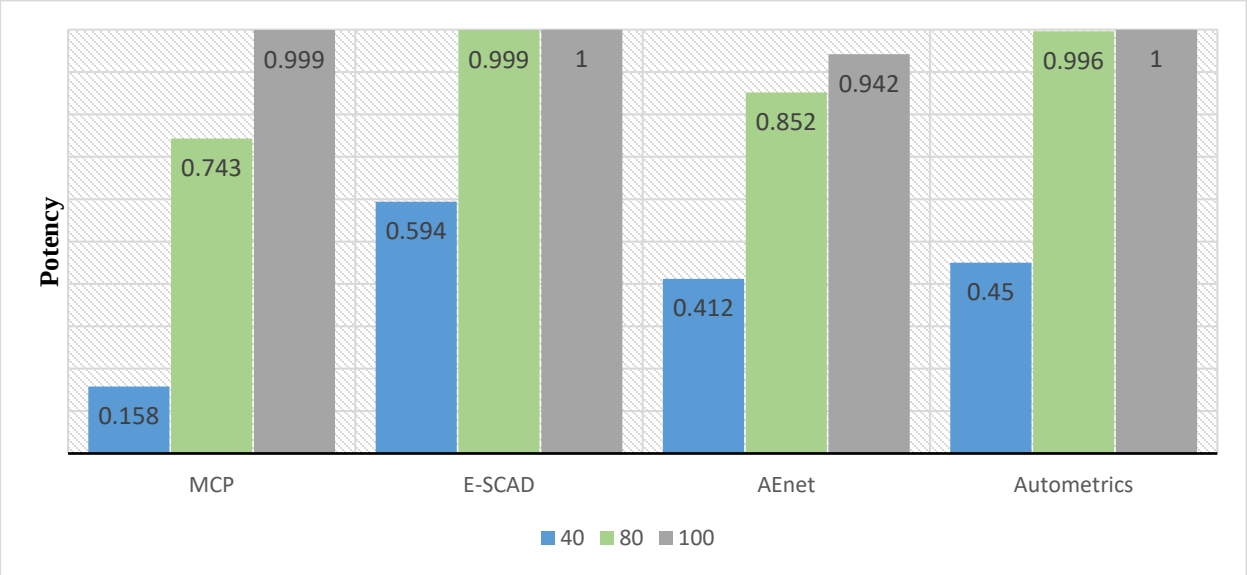
The case of high collinearity among the set of covariates exerts a negative influence on MCP and Autometrics related potency, whereas E-SCAD achieves a substantial improvement in potency. Moreover, the AEnet potency improved when n = 40, but the relative performance (against the moderate case of multicollinearity) did not improve under n = 80 and 100. Increasing the length of covariates (relevant and irrelevant) enhances the potency of all methods. Moreover, the AEnet beats its rival counterparts in gauge. Across the three levels of multicollinearity, it is noted that by switching from $\sum = 0.25$ to $\sum = 0.50$ ($\sum$ shows multicollinearity), the potency of all methods improved, except for Autometrics. Only E-SCAD improved in potency when we changed to $\sum = 0.90$, as shown in Fig. 4.6(a), but the gauge of autometrics is close to zero in Fig. 4.6(b).

Table 4.4. Variable Selection under Multicollinearity from Monte Carlo Simulation

Models	$\sum$ ***= 0.25, P = 130***		$\sum$ ***= 0.25, P = 150***	
n= 40/80/100	**Potency**	**Gauge**	**Potency**	**Gauge**
MCP	0.109/0.461/0.898	0.014/0.054/0.049	0.158/0.743/0.999	0.019/0.062/0.030
E-SCAD	0.507/0.971/1	0.143/0.105/0.083	0.594/0.999/1	0.146/0.099/0.094
AEnet	0.516/0.915/0.971	0.012/0.0006/0	0.412/0.852/0.942	0.016/0.0004/0
Autometrics	0.558/0.988/0.999	0.0004/0/0	0.450/0.996/1	0.0005/0/0
n= 40/80/100	$\sum$ ***= 0.50, P = 130***		$\sum$ ***= 0.50, P = 150***	
MCP	0.189/0.524/0.726	0.025/0.057/0.051	0.255/0.651/0.906	0.034/0.056/0.028
E-SCAD	0.688/0.985/0.999	0.121/0.059/0.031	0.757/0.998/0.999	0.111/0.045/0.030
AEnet	0.409/0.859/0.947	0.015/0.0004/0	0.513/0.919/0.974	0.011/0.0002/0
Autometrics	0.426/0.910/0.993	0.0002/0/0	0.448/0.954/0.997	0.0006/0/0
n= 40/80/100	$\sum$ ***= 0.90, P = 130***		$\sum$ ***= 0.90, P = 150***	
MCP	0.186/0.23/0.236	0.014/0.005/0.003	0.205/0.244/0.253	0.013/0.004/0.002
E-SCAD	0.999/0.999/1	0.03/0.018/0.017	0.999/1/1	0.028/0.020/0.019
AEnet	0.627/0.761/0.796	0/0/0	0.687/0.808/0.840	0/0/0
Autometrics	0.359/0.476/0.524	0/0.0001/0	0.368/0.495/0.545	0.001/0/0

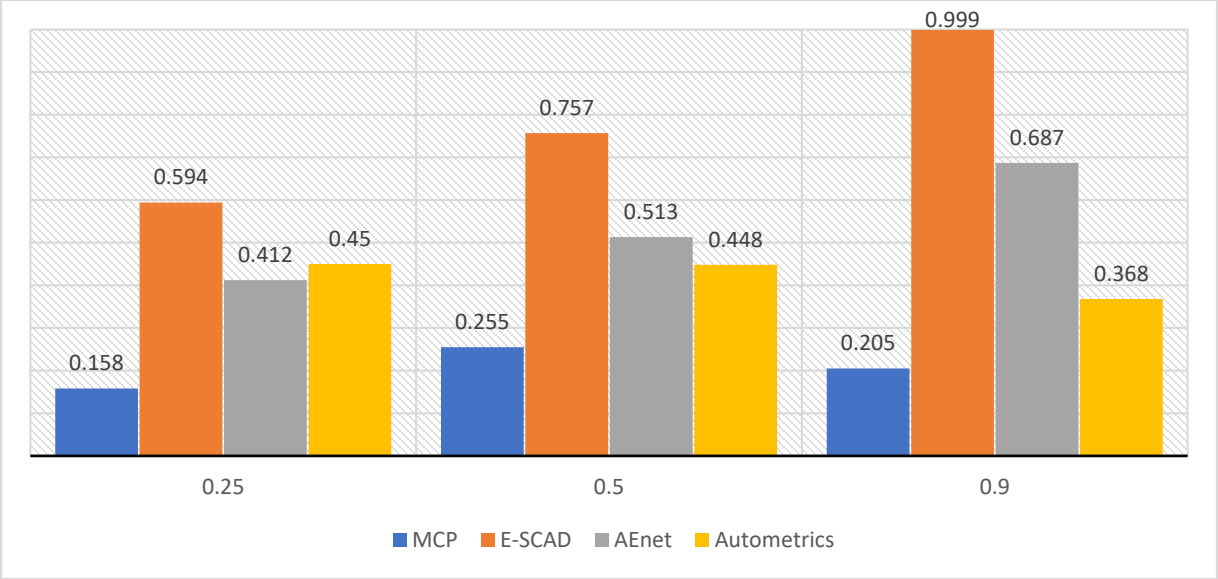


(a)

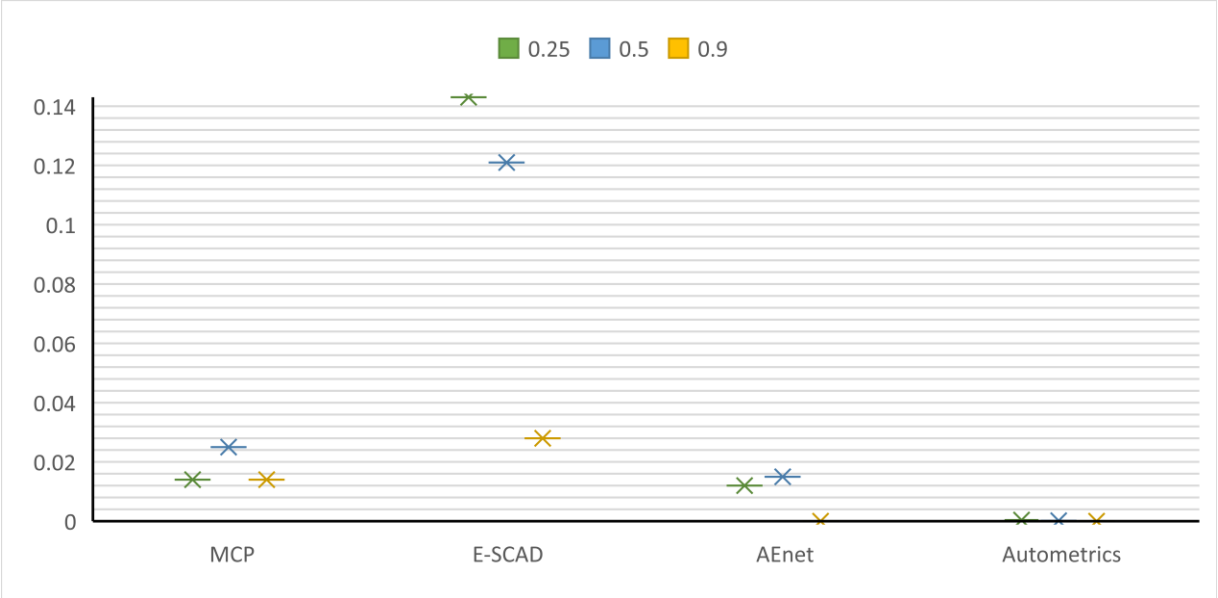


(b)

Figure 4.5: Computation of Potency across sample sizes when $\sum$ = 0.25, P = 130(a) and P = 150(b).



(a)



(b)

Figure 4.6: Computation of Potency and gauge across low, moderate, and high levels of multicollinearity when n = 40 and P = 130(a), P=150(b)

4.2.3. SCENARIO-II

Table 4.5 presents the simulation results by varying heteroscedasticity along with sample size and many covariates (both relevant and irrelevant). The potency of all techniques is growing with the expansion of the data window, as portrayed in Fig. 4.7. In case of low heteroscedastic errors, it can be seen that the potency of Autometrics is close to one if the number of observations is 40, but by increasing the number of observations to 100, both E-SCAD and Autometrics retain all the relevant variables. As we consider more candidate variables (both relevant and irrelevant), which in turn negatively influence the potency of all methods, particularly in case of n = 40 and 80. In contrast, Autometrics shows remarkable performance in gauge whatever the number of predictors (P = 130 or 150). For n = 80 and 100, the AEnet also reduces the gauge to zero.

Increasing the level of heteroscedasticity exerts a huge negative influence on the potency of Autometrics, and more specifically, under the number of 150 predictors, the E-SCAD yields good output, as shown by Fig. 4.8. In terms of gauge, the Autometrics performed very well. Fig. 4.9(a, b) shows that the gauge associated with Autometrics is lower than that of competitor methods.

Table 4.5. 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 = 40/80/100	Potency	Gauge	Potency	Gauge
MCP	0.259/0.781/0.983	0.034/0.034/0.002	0.196/0.567/0.851	0.025/0.055/0.022
E-SCAD	0.776/0.999/1	0.102/0.002/0.0001	0.692/0.995/0.999	0.116/0.015/0.002
AEnet	0.543/0.995/0.999	0.009/0/0	0.428/0.967/0.999	0.014/0/0
Autometrics	0.999/1/1	0/0/0	0.975/1/1	0/0/0
n = 40/80/100	π_2 **= 0.2/0.6,** ***P = 130***		π_2 **= 0.2/0.6,** ***P = 150***	
MCP	0.261/0.759/0.98	0.035/0.037/0.003	0.195/0.563/0.834	0.026/0.054/0.026
E-SCAD	0.776/0.999/1	0.104/0.004/0.001	0.692/0.994/0.999	0.116/0.019/0.003
AEnet	0.540/0.989/0.999	0.009/0/0	0.435/0.953/0.997	0.032/0/0
Autometrics	0.924/1/1	0/0/0	0.645/1/1	0.0001/0/0
n = 40/80/100	π_3 **= 0.3/0.9,** ***P = 130***		π_3 **= 0.3/0.9,** ***P = 150***	
MCP	0.257/0.729/0.972	0.034/0.043/0.005	0.192/0.549/0.812	0.025/0.055/0.031
E-SCAD	0.770/0.999/1	0.105/0.009/0.003	0.694/0.992/0.999	0.118/0.028/0.007
AEnet	0.532/0.975/0.997	0.010/0.00003/0	0.421/0.929/0.989	0.015/0/0
Autometrics	0.657/1/1	0.0003/0/0	0.518/0.999/1	0.0002/0/0

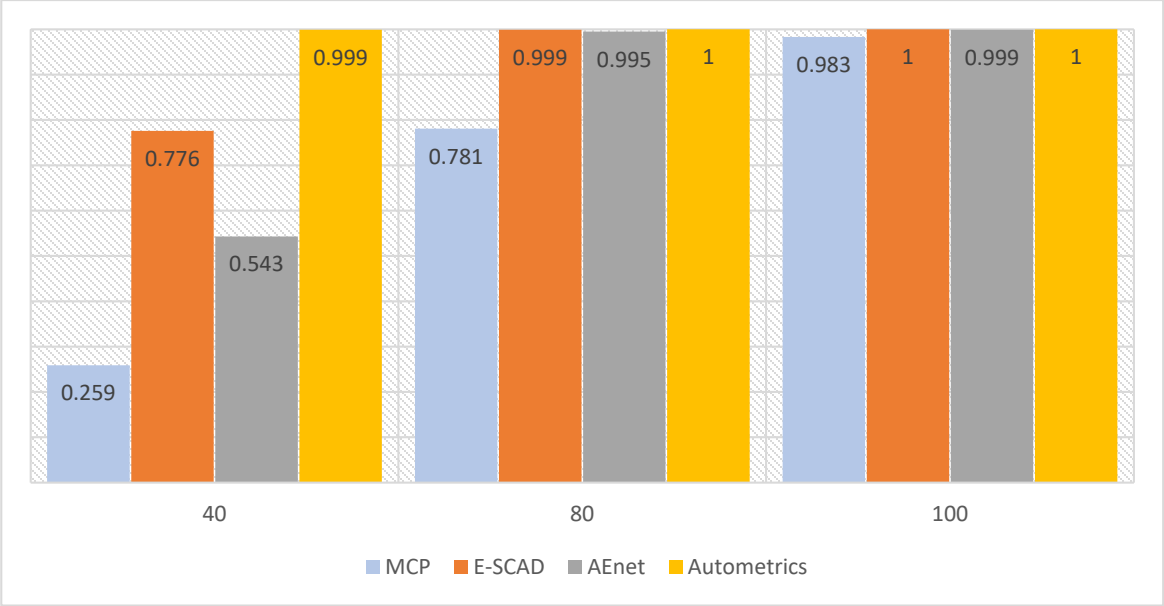


Figure 4.7: Computation of Potency across sample sizes when $\boldsymbol{\pi_1}$ = 0.1/0.3, P = 130

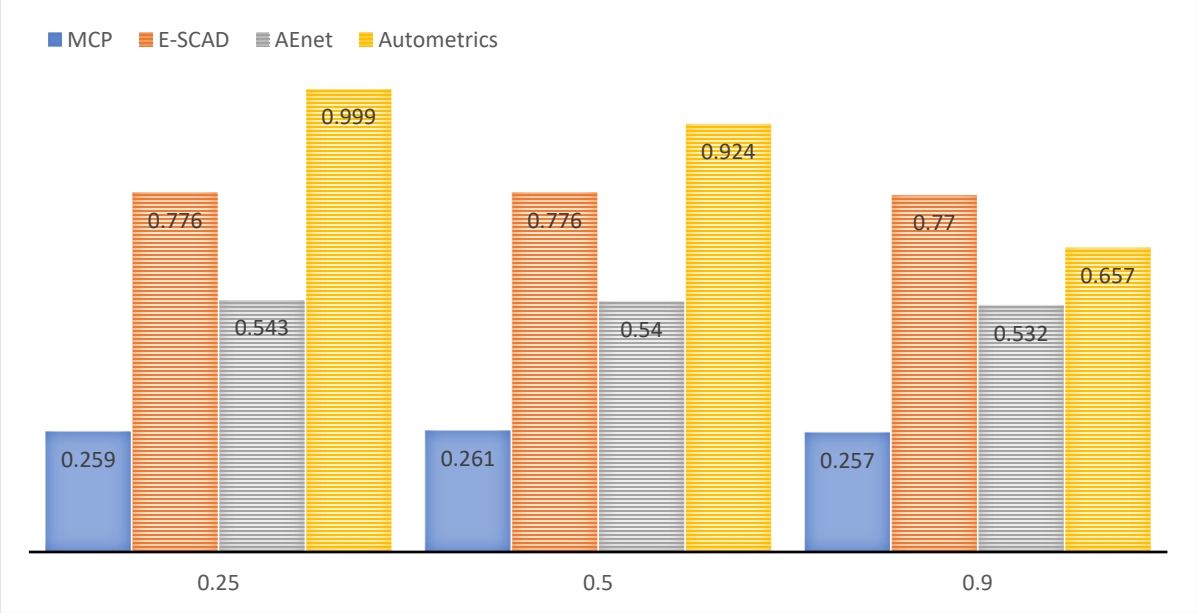
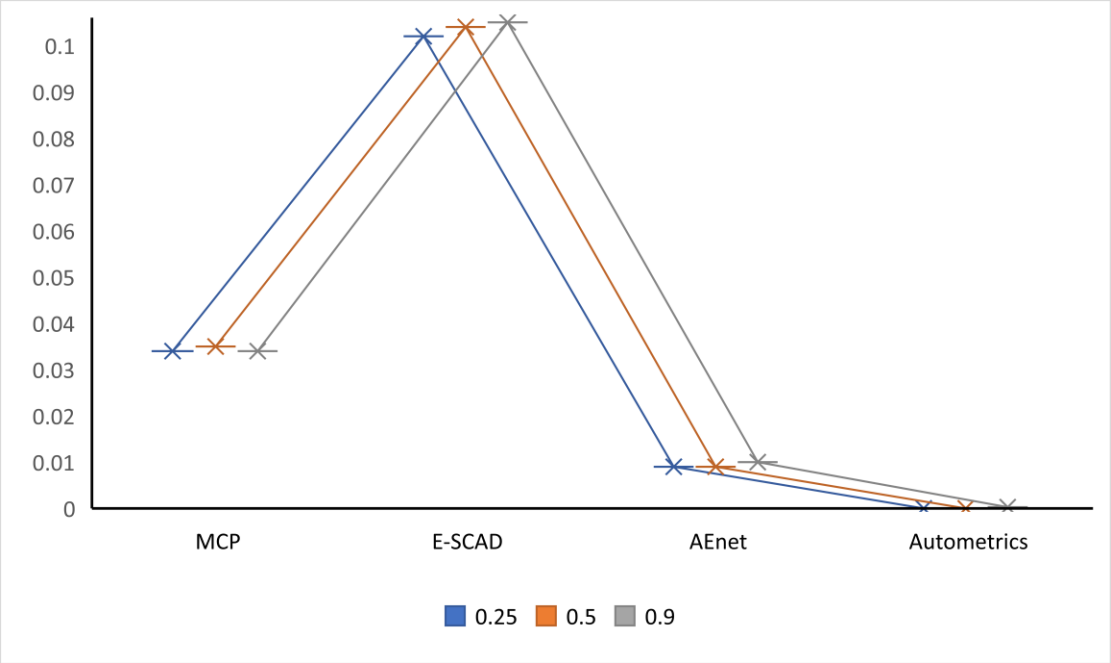
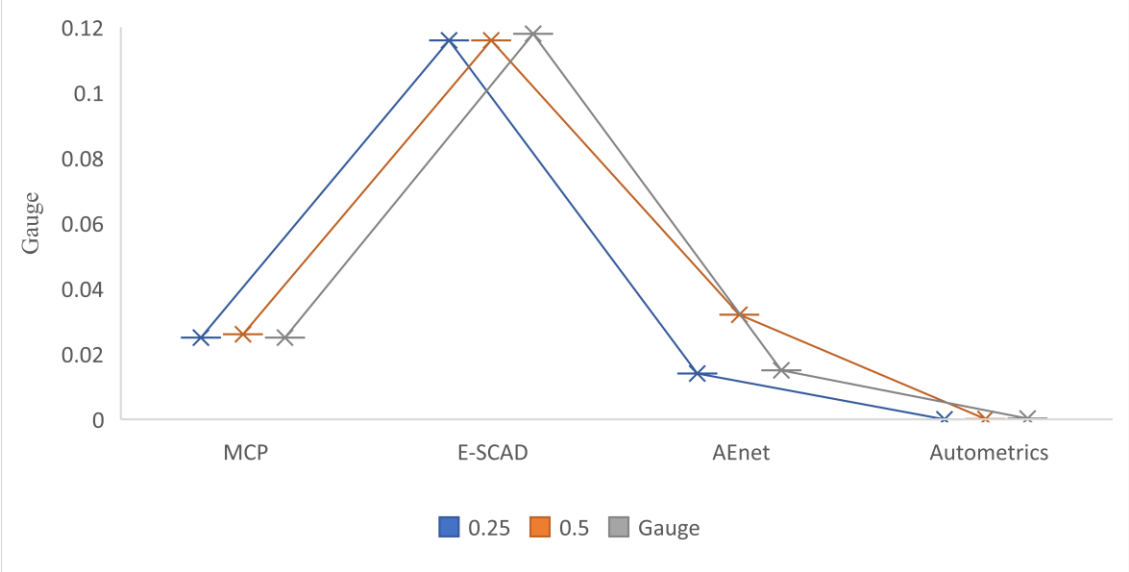


Figure 4.8: Computation of Potency across all levels of Heteroscedasticity when n = 40 and P = 130.



(a)



(b)

Figure 4.9: Computation of Potency across all levels of Heteroscedasticity when n = 40, P = 130 and P=150.

4.2.4. SCENARIO-III

Whatever the level of Autocorrelation**,** the potency of E-SCAD is often higher than the competitive counterparts for a different number of covariates we use (P = 130/150), as shown in Fig. 4.10. It can be observed in Fig. 4.10 that the potency of Autometrics rapidly tends to one asymptotically. Moreover, it can be seen in Fig. 4.11 that all methods' performance gets worse, as we raise the level of Autocorrelation. In terms of gauge, the Autometrics showed good performance, which circumvents the inclusion of irrelevant variables. The AEnet is a good competitor to Autometrics in gauge, particularly when n = 80 and 100.

Table 4.6. Variable selection under Autocorrelation from Monte Carlo Simulation

Models	***ρ* = 0.25, *P* = 130**		***ρ* = 0.25, *P* = 150**	
n = 40/80/100	**Potency**	**Gauge**	**Potency**	**Gauge**
MCP	0.245/0.648/0.903	0.032/0.057/0.031	0.186/0.522/0.717	0.025/0.057/0.053
E-SCAD	0.750/0.997/0.999	0.115/0.050/0.034	0.689/0.984/0.999	0.122/0.060/0.035
AEnet	0.516/0.915/0.971	0.118/0.0001/0	0.410/0.850/0.941	0.014/0.0004/0
Autometrics	0.446/0.936/1	0.001/0/0	0.441/0.887/0.987	0.0001/0/0
n = 40/80/100	***ρ* = 0.50, *P* = 130**		***ρ* = 0.50, *P* = 150**	
MCP	0.250/0.628/0.868	0.032/0.058/0.041	0.183/0.513/0.688	0.025/0.059/0.057
E-SCAD	0.752/0.996/0.999	0.118/0.063/0.048	0.690/0.980/0.998	0.120/0.071/0.046
AEnet	0.514/0.900/0.959	0.009/0.0002/0	0.406/0.832/0.922	0.015/0.0006/0
Autometrics	0.427/0.866/0.976	0.001/0/0	0.408/0.806/0.956	0.0002/0/0
n = 40/80/100	***ρ* = 0.90, *P* = 130**		***ρ* = 0.90, *P* = 150**	
MCP	0.234/0.524/0.634	0.030/0.061/0.068	0.174/0.450/0.562	0.023/0.058/0.065
E-SCAD	0.730/0.952/0.982	0.131/0.145/0.148	0.671/0.926/0.970	0.132/0.133/0.134
AEnet	0.459/0.766/0.829	0.016/0.002/0.001	0.337/0.701/0.781	0.017/0.003/0.001
Autometrics	0.321/0.503/0.573	0.001/0.0002/0.0002	0.315/0.474/0.542	0.0002/0.0001/0

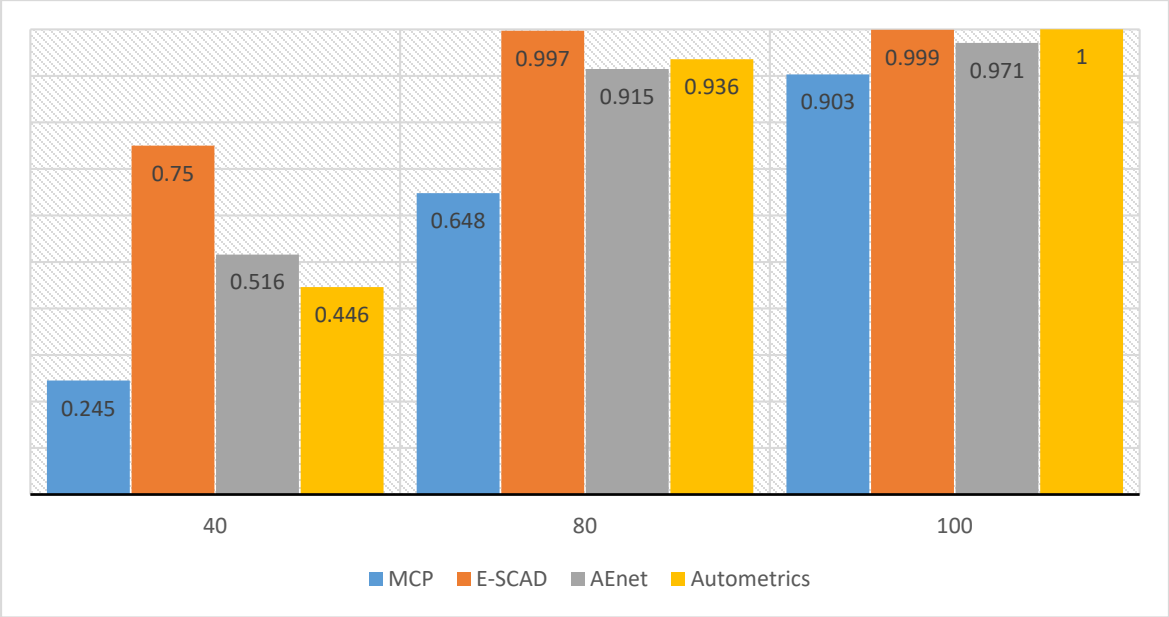


Figure 4.10: Computation of Potency across the sample sizes when ρ = 0.25, P = 130.

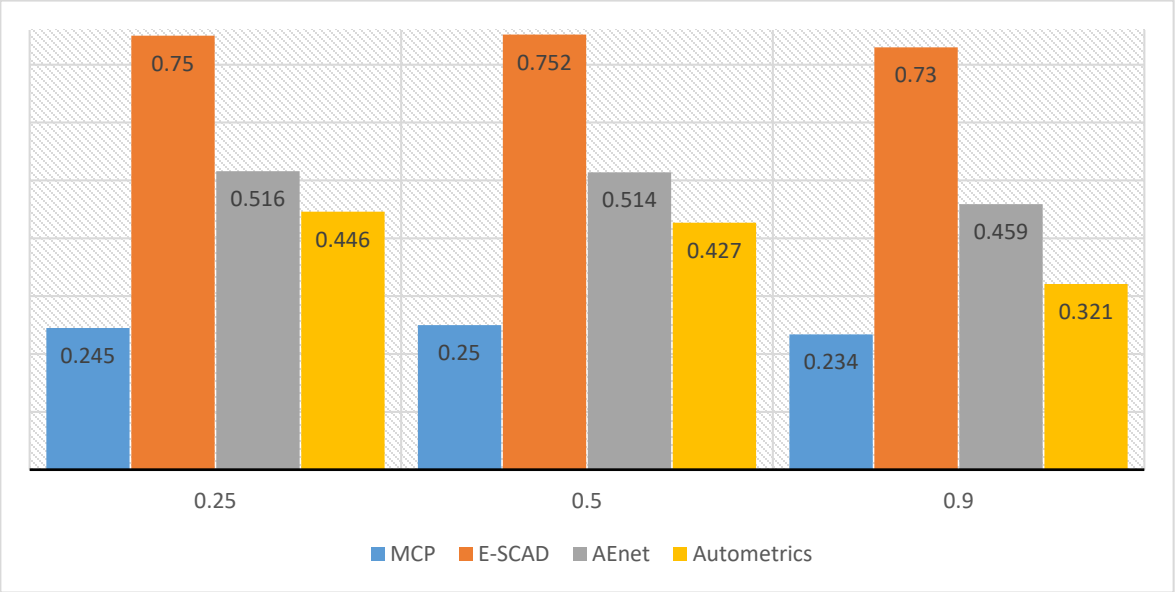


Figure 4.11: Computation of Potency across all levels of Autocorrelation when n = 40, P = 130.

Chapter 5

Forecast Comparison and Discussion

5.1. Out-of-Sample Forecasting Comparison using Huge Big Data

Prediction is a very difficult art, especially when it involves the future‖ -Neils Bohr ***(Nobel Laureate Physicist)***.

The prediction of macroeconomic variables is very important in macroeconomic studies, monetary policy analysis, and environmental economics. Accurate forecasts lead to a better understanding of dynamic economy mechanisms (Bai and Ng, 2008), more effective monetary policies (Bernanke et al. 2005), and improved portfolio management and hedging strategies (Rapach et al. 2010). In the data-rich environment of today, economists and those who make decisions keep an eye on many macroeconomic series.

Low-dimensional models often include some pre-specified economic covariates, for instance, vector auto-regression, and therefore have a complication in capturing the dynamic and complex patterns that contain huge panels of time series (Li and Chen, 2014). An under-specified model produces biased results when important variables are missing. There is a strong need to propose updated statistical models and analytical frameworks with the goal of expanding the low-dimensional counterparts to make better predictions.

5.1.1. Simulation Results

This section uses the same design of experiments, i.e., the number of observations and the number of variables (p and q), which were explained in detail in section 4.1.

The forecast comparison results derived from Monte Carlo experiments are presented in Tables 5.1-5.3. All methods are improving their performance by augmenting the number of observations,

as shown in Fig. 5.1. On the other hand, when we add more irrelevant variables, the methods become less good at making predictions.

5.1.1.1. SCENARIO-I

In presence of low multicollinearity, it can be observed in Fig. 2(a) (see Appendix B) that the forecasting performance of MCP is superior to other rival methods across different sample sizes when the number of predictors is 50. Although we expand the variable window to 70, the MCP remains dominant except for the small sample case, where E-SCAD outperforms the competitive methods. Increasing the level of multicollinearity among variables, E-SCAD produced a better forecast for a small sample size, but as the sample size increased, the MCP produced a more satisfactory forecast than its competitor counterparts, as revealed by Fig. 2(b) (see Appendix B). Considering the case of extreme collinearity, it can be seen from Fig. 2(c) (see Appendix B) that the PLS-based factor model is superior, in particular at n = 80, while asymptotically, E-SCAD outperformed its rival counterparts. Furthermore, as shown in Fig. 5.2(a and b), factor models improved in forecasting accuracy by increasing the levels of multicollinearity. Similarly, the E-SCAD showed improvement in accuracy in Fig. 5.2(b).

Table 5.1. Forecast Comparison under Multicollinearity from Monte Carlo Simulation

Models	$\sum$ = ***0.25, P = 50***		$\sum$ = ***0.25, P = 70***	
n=80/160/320	**RMSE**	**MAE**	**RMSE**	**MAE**
MCP	**1.123/1.055/1.027**	**0.908/0.848/0.821**	1.205/**1.069/1.031**	0.971/**0.858/0.825**
E-SCAD	1.135/1.066/1.034	0.917/0.856/0.827	**1.195**/1.086/1.040	**0.961**/0.872/0.831
AEnet	1.237/1.131/1.070	0.996/0.911/0.856	1.304/1.152/1.083	1.058/0.925/0.867
Autometrics	1.316/1.091/1.027	1.065/0.874/0.822	1.316/1.091/1.042	1.065/0.874/0.834
FM_PCA	3.517/3.210/2.829	2.839/2.576/2.260	4.493/4.305/3.966	3.623/3.458/3.173
FM_PLS	1.528/1.200/1.090	1.235/0.963/0.871	1.921/1.321/1.126	1.551/1.059/0.901
n=80/160/320	$\sum$ = ***0.5, P = 50***		$\sum$ = ***0.5, P = 70***	
MCP	1.145/**1.056/1.027**	0.925/**0.848/0.821**	1.318/**1.069/1.032**	1.062/**0.858/0.825**
E-SCAD	**1.112**/1.058/1.030	**0.898**/0.849/0.824	**1.168**/1.074/1.035	**0.940**/0.862/0.827
AEnet	1.282/1.147/1.077	1.031/0.924/0.862	1.341/1.176/1.093	1.088/0.943/0.874
Autometrics	1.156/1.062/1.027	0.931/0.853/0.821	1.473/1.091/1.041	1.191/0.874/0.833
FM_PCA	2.583/2.053/1.705	2.088/1.644/1.365	3.933/3.334/2.700	3.174/2.677/2.164
FM_PLS	1.368/1.161/1.080	1.105/0.932/0.864	1.595/1.248/1.108	1.287/1.001/0.886
n=80/160/320	$\sum$ = ***0.9, P = 50***		$\sum$ = ***0.9, P = 70***	
MCP	1.484/1.157/1.042	1.198/0.930/0.832	1.764/1.261/1.058	1.424/1.013/0.846
E-SCAD	1.201/**1.060/1.019**	0.968/**0.851/0.814**	1.291/**1.080/1.021**	1.040/**0.867/0.817**
AEnet	1.327/1.180/1.099	1.067/0.950/0.879	1.422/1.227/1.129	1.152/0.985/0.903
Autometrics	4.363/1.795/1.031	3.528/1.443/0.825	6.589/2.501/1.053	5.333/2.006/0.843
FM_PCA	1.169/1.099/1.075	0.943/0.883/0.859	1.318/1.212/1.165	1.065/0.974/0.932
FM_PLS	**1.138**/1.078/1.043	**0.919**/0.865/0.834	**1.184**/1.095/1.053	**0.959**/0.880/0.842

Noted: Bold values indicate a better forecast.

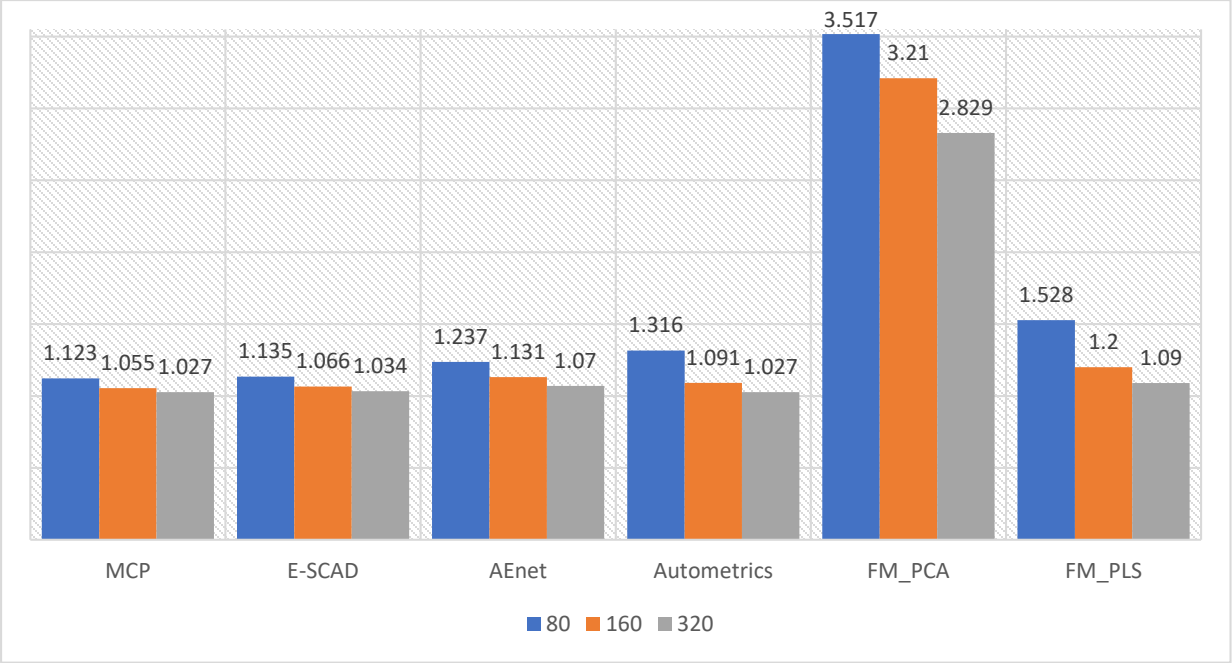
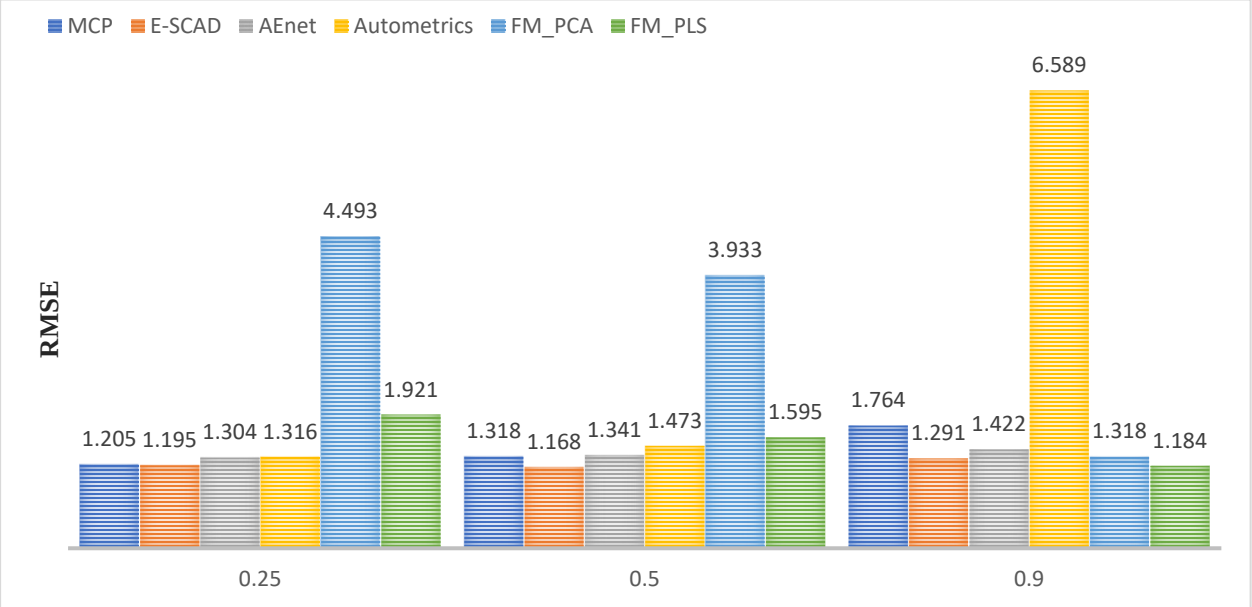
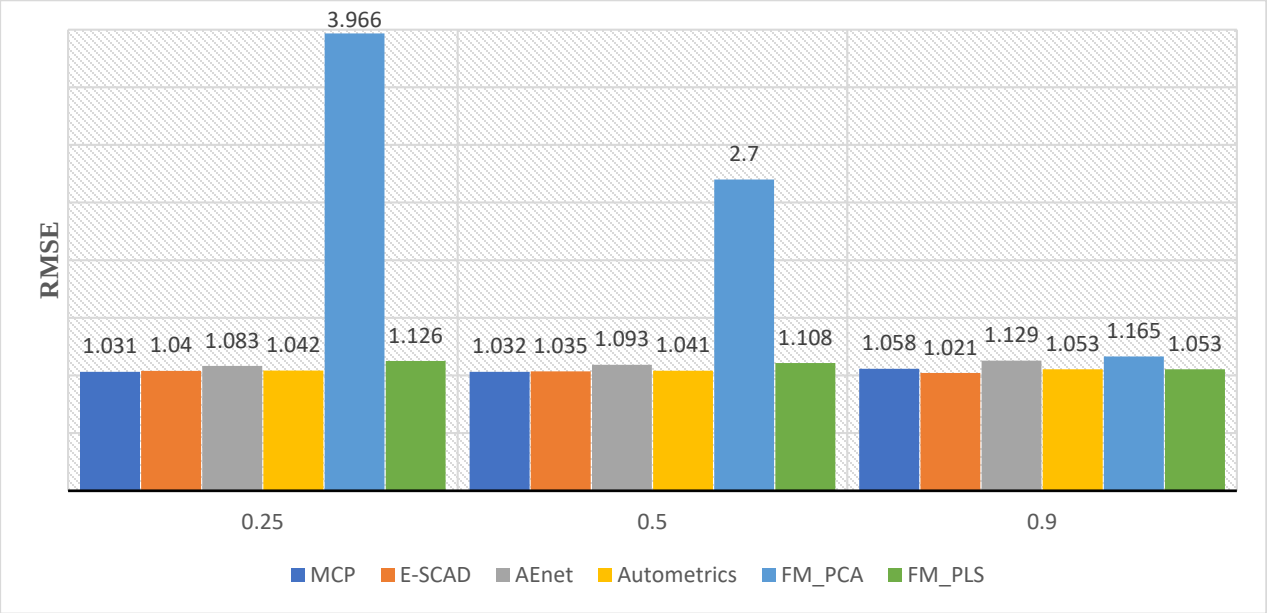


Figure 5.1: Out of sample root mean squares error across sample sizes, where forecasts are obtained from various models when ρ = 0.25(a) and P = 50.



(a)



(b)

Figure 5.2: Out of sample root mean squares error across the levels of Multicollinearity, where forecasts are obtained from various models when n = 80(a), n = 320(b), and P = 70.

5.1.1.2. SCENARIO-II

In presence of all schemes of heteroscedasticity, the performance of MCP is often better than all of its competitor counterparts. Whatever the number of predictors to be used, whether 50/70, the accuracy of the MCP forecast is maintained and is dominant all the time. Apart from this, when the number of predictors is equal to 50, Autometrics provides a similar forecast as MCP in large samples. In addition, Fig. 3(a, b and c) (see Appendix B) shows the improvement in the accuracy level with expanding the training data window, whatever the values of σ we consider here, i.e., 0.2/0.6 or 0.3/0.9. Although switching from 0.2/0.6 to 0.3/0.9, the RMSE associated with each tool tends to increase. This increase was also observed when the triplet size of heteroscedastic errors was considered, as well as when the candidate set variables (relevant and irrelevant) were varied in Fig. 4(a and b) (see Appendix B).

Table 5.2. Forecast Comparison under Heteroscedasticity from Monte Carlo Simulation

Models	$\boldsymbol{\pi_1}$ **= 0.1/0.3,** ***P = 50***		$\boldsymbol{\pi_1}$ **= 0.1/0.3,** ***P = 70***	
n=80/160/320	**RMSE**	**MAE**	**RMSE**	**MAE**
MCP	**0.313/0.306/0.303**	**0.253/0.246/0.242**	**0.321/0.307/0.303**	**0.260/0.246/0.242**
E-SCAD	0.319/0.309/0.304	0.258/0.248/0.243	0.331/0.311/0.305	0.267/0.249/0.243
AEnet	0.331/0.318/0.313	0.268/0.255/0.250	0.354/0.326/0.314	0.286/0.262/0.251
Autometrics	0.318/0.308/**0.303**	0.256/0.248/**0.242**	0.339/0.313/0.305	0.274/0.250/0.244
FM_PCA	3.373/3.055/2.648	2.723/2.452/2.115	4.382/4.197/3.847	3.534/3.374/3.078
FM_PLS	0.399/0.327/0.311	0.322/0.262/0.249	0.625/0.347/0.317	0.504/0.278/0.253
n=80/160/320	$\boldsymbol{\pi_2}$ **= 0.2/0.6,** ***P = 50***		$\boldsymbol{\pi_2}$ **= 0.2/0.6,** ***P = 70***	
MCP	**0.627/0.613/0.606**	**0.507/0.492/0.484**	**0.643/0.614/0.607**	**0.520/0.492/0.485**
E-SCAD	0.637/0.617/0.609	0.515/0.496/0.486	0.659/0.621/0.609	0.532/0.498/0.487
AEnet	0.662/0.636/0.623	0.537/0.510/0.499	0.683/0.644/0.625	0.550/0.518/0.500
Autometrics	0.636/0.617/**0.606**	0.512/0.496/**0.484**	0.667/0.625/0.610	0.548/0.501/0.488
FM_PCA	3.410/3.101/2.704	2.753/2.489/2.160	4.412/4.233/3.883	3.556/3.402/3.106
FM_PLS	0.798/0.654/0.623	0.646/0.525/0.498	1.107/0.693/0.634	0.892/0.556/0.507
n=80/160/320	$\boldsymbol{\pi_3}$ **= 0.3/0.9,** ***P = 50***		$\boldsymbol{\pi_3}$ **= 0.3/0.9,** ***P = 70***	
MCP	**0.941/0.920/0.909**	**0.761/0.739/0.727**	**0.965/0.921/0.910**	**0.780/0.739/0.728**
E-SCAD	0.954/0.926/0.913	0.771/0.743/0.730	0.985/0.930/0.914	0.795/0.746/0.730
AEnet	1/0.956/0.936	0.810/0.768/0.749	1.027/0.969/0.939	0.828/0.779/0.751
Autometrics	0.954/0.926/**0.909**	0.768/0.744/**0.727**	1.017/0.938/0.916	0.823/0.752/0.733
FM_PCA	3.478/3.176/2.791	2.809/2.549/2.230	4.467/4.281/3.941	3.601/3.440/3.153
FM_PLS	1.181/0.983/0.935	0.956/0.789/0.748	1.507/1.040/0.951	1.215/0.834/0.760

Noted: Bold values indicate a better forecast.

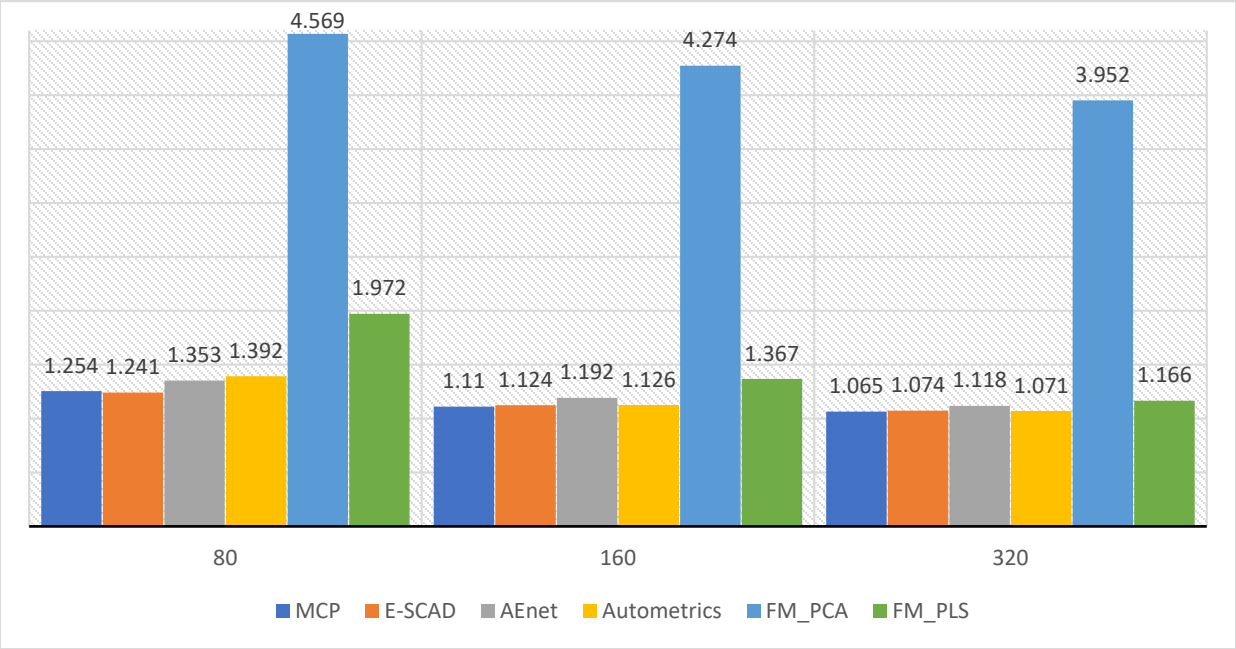
5.1.1.3. SCENARIO-III

In presence of low and moderate autocorrelation, the MCP shows outstanding performance in terms of forecasting, particularly when we increase the sample size, as shown by Fig. 5.3(a, b). In contrast, when n = 80, the E-SCAD produced a remarkable forecast. In the case of extreme autocorrelation, E-SCAD outperformed the rival techniques under both 80 and 160 data points, but as we further augmented the sample to 320, the MCP induced a more accurate forecast. All these results are supported by Fig. 5.3(c). Furthermore, it can be noticed from Fig. 5.4(a and b) that the predictive ability of all tools distorts with increasing the level of Autocorrelation. As we progress from moderate autocorrelation to extreme autocorrelation, the Autometrics efficacy is more negatively affected than the MCP, E-SCAD, and PLS-based factor models, as shown in Fig. 5.4(a).

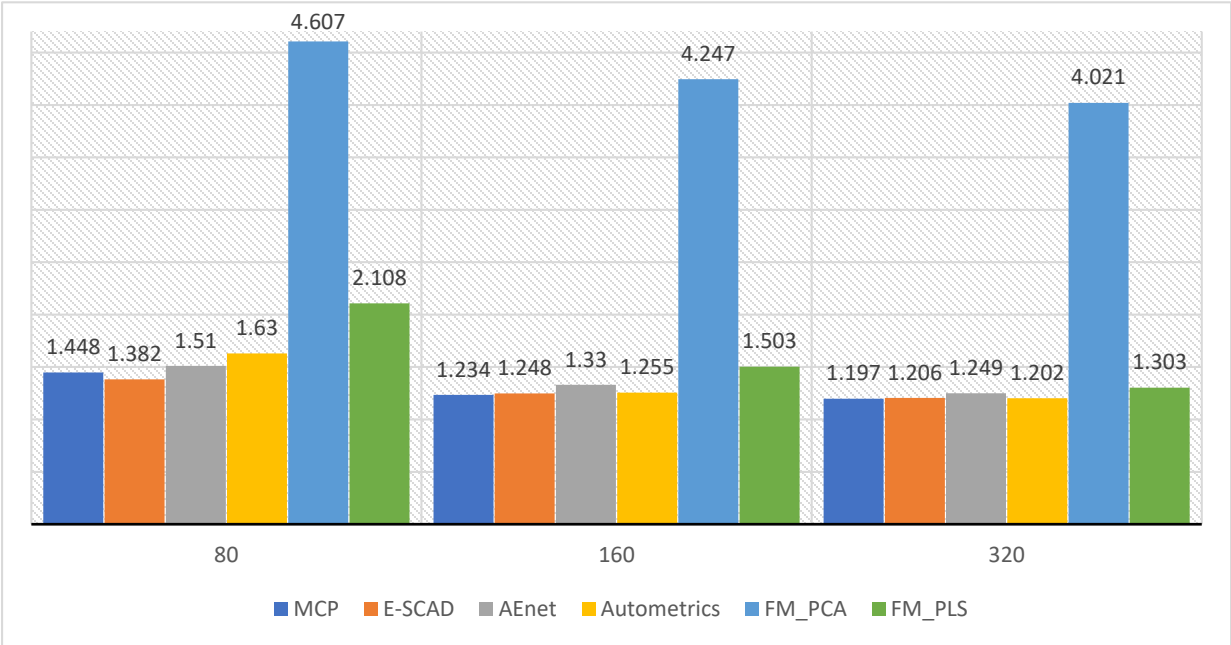
Table 5.3. Forecast comparison under Autocorrelation from Monte Carlo Simulation

Models	***ρ* = 0.25, *P* = 50**		***ρ* = 0.25, *P* = 70**	
n=80/160/320	**RMSE**	**MAE**	**RMSE**	**MAE**
MCP	**1.167/1.078/1.056**	**0.943/0.866/0.845**	1.254/**1.110/1.065**	1.012/**0.892/0.851**
E-SCAD	1.175/1.091/1.062	0.952/0.877/0.850	**1.241**/1.124/1.074	**1.002**/0.904/0.859
AEnet	1.250/1.174/1.104	1.012/0.944/0.886	1.353/1.192/1.118	1.094/0.957/0.895
Autometrics	1.192/1.100/1.064	0.963/0.884/0.851	1.392/1.126/1.071	1.121/0.908/0.858
FM_PCA	3.520/3.222/2.858	2.848/2.589/2.288	4.569/4.274/3.952	3.695/3.429/3.165
FM_PLS	1.568/1.231/1.119	1.268/0.990/0.896	1.972/1.367/1.166	1.591/1.101/0.932
n=80/160/320	***ρ* = 0.50, *P* = 50**		***ρ* = 0.50, *P* = 70**	
MCP	1.324/**1.222/1.185**	1.073/**0.987/0.949**	1.448/**1.234/1.197**	1.177/**0.993/0.957**
E-SCAD	**1.318**/1.238/1.191	**1.068**/0.996/0.954	**1.382**/1.248/1.206	**1.122**/1.005/0.965
AEnet	1.409/1.310/1.230	1.140/1.056/0.985	1.510/1.33/1.249	1.225/1.070/1.001
Autometrics	1.330/1.222/1.187	1.080/0.985/0.951	1.630/1.255/1.202	1.318/1.011/0.964
FM_PCA	3.570/3.279/2.916	2.889/2.624/2.333	4.607/4.247/4.021	3.716/3.381/3.219
FM_PLS	1.720/1.392/1.258	1.389/1.121/1.005	2.108/1.503/1.303	1.702/1.206/1.042
n=80/160/320	***ρ* = 0.90, *P* = 50**		***ρ* = 0.90, *P* = 70**	
MCP	2.953/2.408/**2.364**	2.449/1.997/**1.936**	3.608/2.538/**2.368**	2.961/2.100/**1.940**
E-SCAD	**2.714**/**2.380**/2.366	**2.267**/**1.976**/1.937	**3.039**/**2.498**/2.370	**2.525**/**2.069**/1.941
AEnet	2.812/2.560/2.435	2.346/2.117/1.981	3.081/2.568/2.515	2.549/2.116/2.051
Autometrics	3.250/2.480/2.358	2.693/2.049/1.930	4.273/2.594/2.394	3.494/2.146/1.957
FM_PCA	4.165/3.871/3.563	3.387/3.126/2.868	5.051/4.735/4.506	4.111/3.810/3.609
FM_PLS	2.941/2.579/2.476	2.439/2.122/2.020	3.341/2.796/2.544	2.749/2.293/2.072

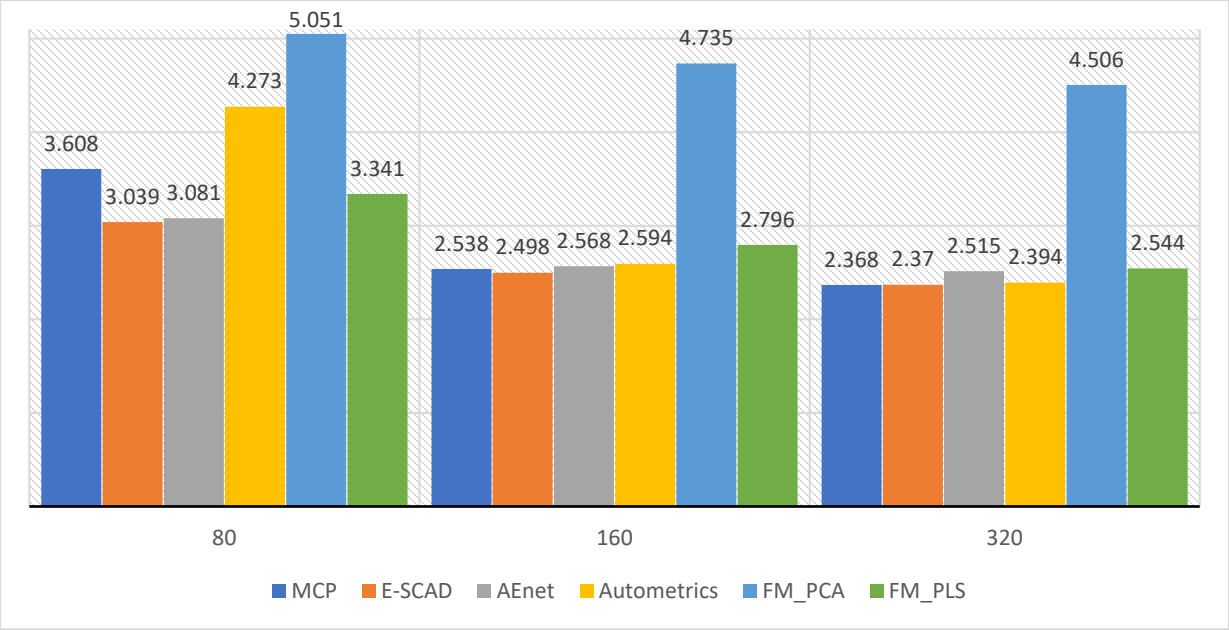
Noted: Bold values indicate a better forecast.


4.569
4.274
3.952
1.972
1.367
1.166
1.254 1.241 1.353 1.392
1.11 1.124 1.192 1.126
1.065 1.074 1.118 1.071
80
160
320
MCP
E-SCAD
AEnet
Autometrics
FM_PCA
FM_PLS

(a)

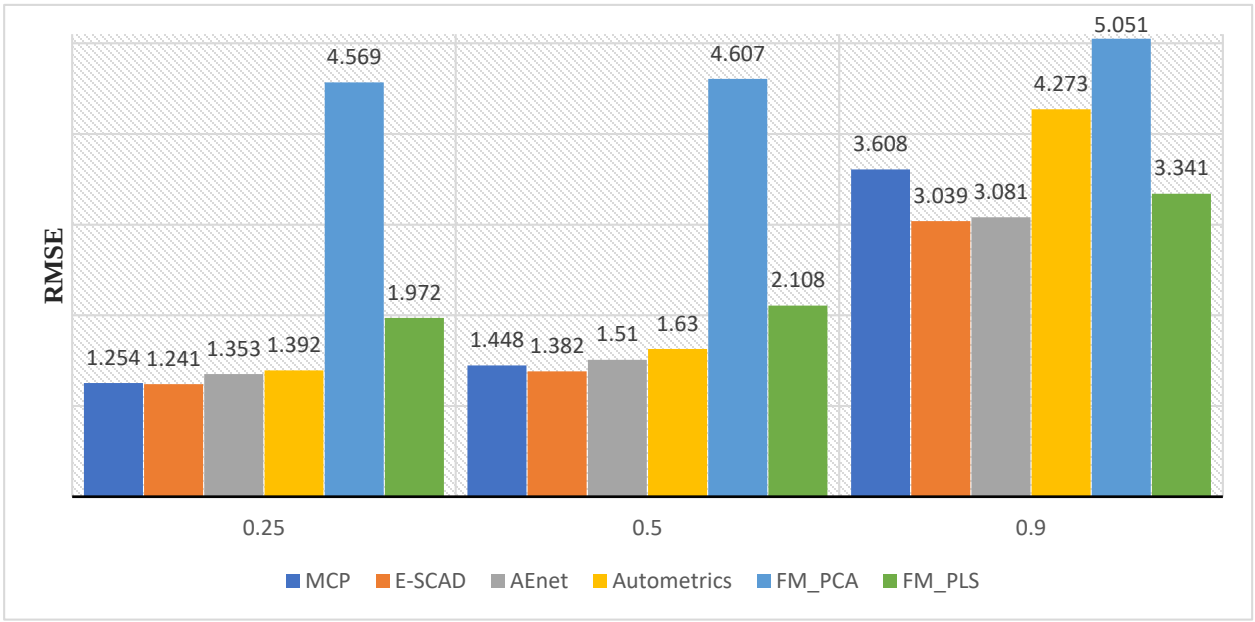

4.607
4.247
4.021
2.108
1.503
1.303
1.448 1.382 1.51 1.63
1.234 1.248 1.33 1.255
1.197 1.206 1.249 1.202
80
160
320
MCP
E-SCAD
AEnet
Autometrics
FM_PCA
FM_PLS

(b)

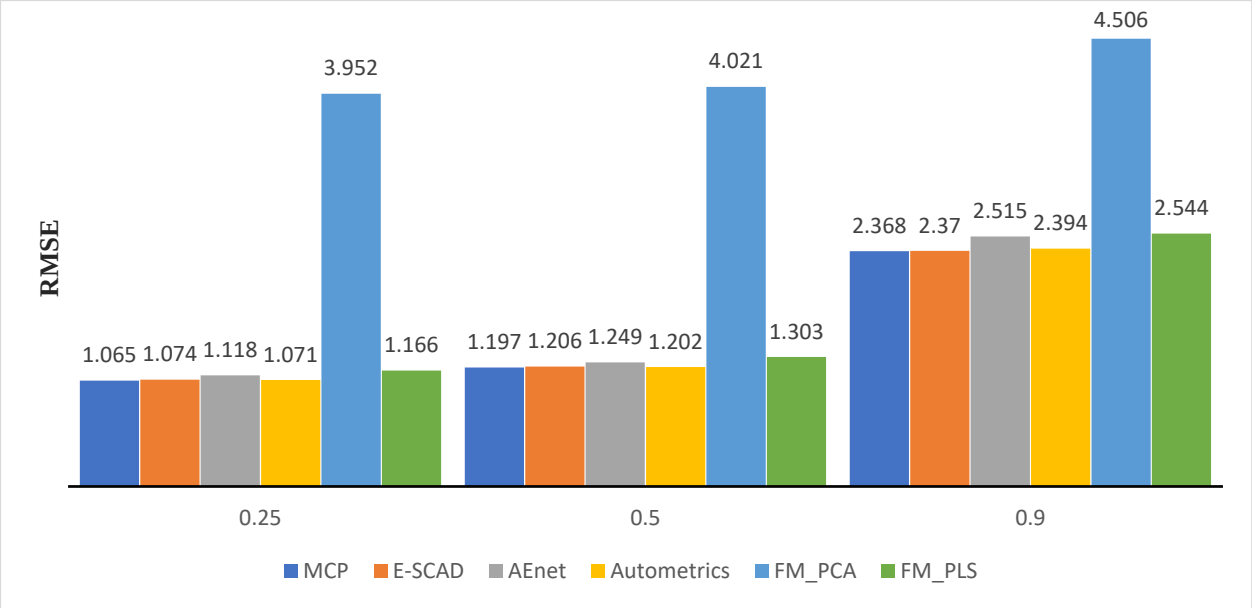


(c)

Figure 5.3: Out of sample root mean squares error across sample size, where forecasts are computed from different models when rho = 025(a), 0.5(b), 0.9(c) and P = 70



(a)



(b)

Figure 5.4: Out of sample root mean squares error across the levels of Autocorrelation, where forecasts are obtained from various models when n = 80(a), 320(b) and P = 70.

5.2. Out-of-Sample Forecasting Comparison using Fat Big Data

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. Albeit, the performance alters depending on the data dimension and estimation tool to be applied as well. Failure in dimensional reduction induces poor output because of redundant variables. Factor models are frequently employed for predictive modelling in a data-rich environment, building on Stock and Watson's (2002a) seminal work on forecasting through diffusion index (DI). Stock and Watson (2012) showed that forecasting via factor models is more accurate than 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 reducing the regression dimension and that it appears to be hard to enhance this performance

without introducing severe changes to the predictive model. Recently, the factor models that are extended for forecasting aim to include those of Hansen and Liao (2019), Bai and Liao (2016), Fan et al. (2016a), Fan et al. (2016b), and Fan et al. (2017).

In addition to the DI methodology, sparse regression is another family of tools utilized for dimension reduction and forecasting and 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 advantage of such tools is that they can deal with the curse of dimensionality, which has been a problem in macroeconomic time series for a long time. However, the predictions that statistical tools make have also been used to make good monetary policies (Bernanke et al., 2005; Syed and Lee, 2020).

5.2.1. Simulation Results

This section uses the same design of experiments, i.e., the number of observations and the number of variables (p and q) as elaborately delineated in Section 4.2.

The forecast comparison output obtained from Monte Carlo exercises is reported in Tables 5.4-5.6. The entries in bold show the best performance of the underlying model.

5.2.1.1. SCENARIO-I

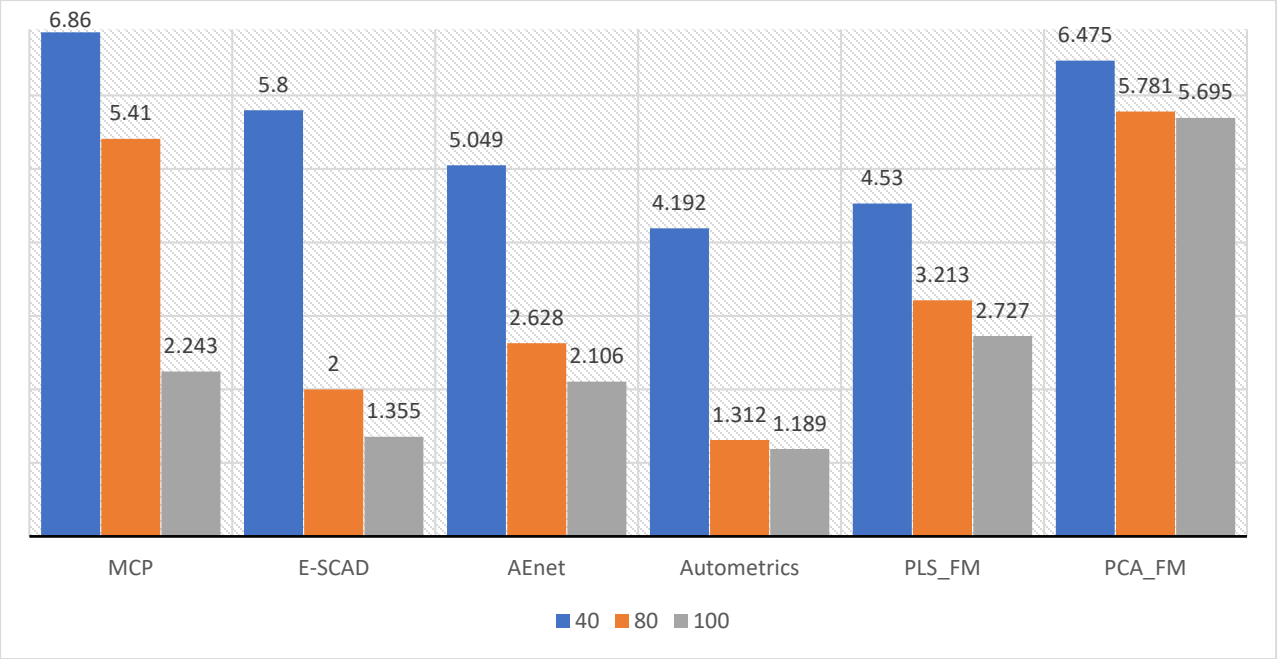
It is observed that the performance of all procedures improves with increasing data points, as shown by Fig. 5.5(a, b, and c). It is also clear from Figure 5.5 that there is a slight impact of sample size on PCA-based factor models. Furthermore, it is clear from Figure 5.6(a, b) that regardless of how large the candidate variables window is and how large the correlation between them is, all methods are effective.

Considering the cases of low and moderate multicollinearity, the forecasting performance of Autometrics is superior to that of its competitive counterparts. But, in the case of a small sample, the RMSE and MAE associated with Autometrics are slightly better than the PLS-based factor approach. It clearly indicates that the PLS-based factor approach is strongly competitive if n is small. Similarly, regardless of the considerable improvement in RMSE and MAE achieved by E-SCAD with increasing the sample size, the forecasting performance is not as satisfactory as Autometrics. Moreover, by increasing the number of relevant and irrelevant variables, Autometrics remains dominant with the lowest RMSE and MAE. In presence of extreme multicollinearity, the factor approach based on PLS outperformed its rival counterparts in terms of the lowest forecast error. Although, according to both error criteria, Autometrics stood as a good competitor.

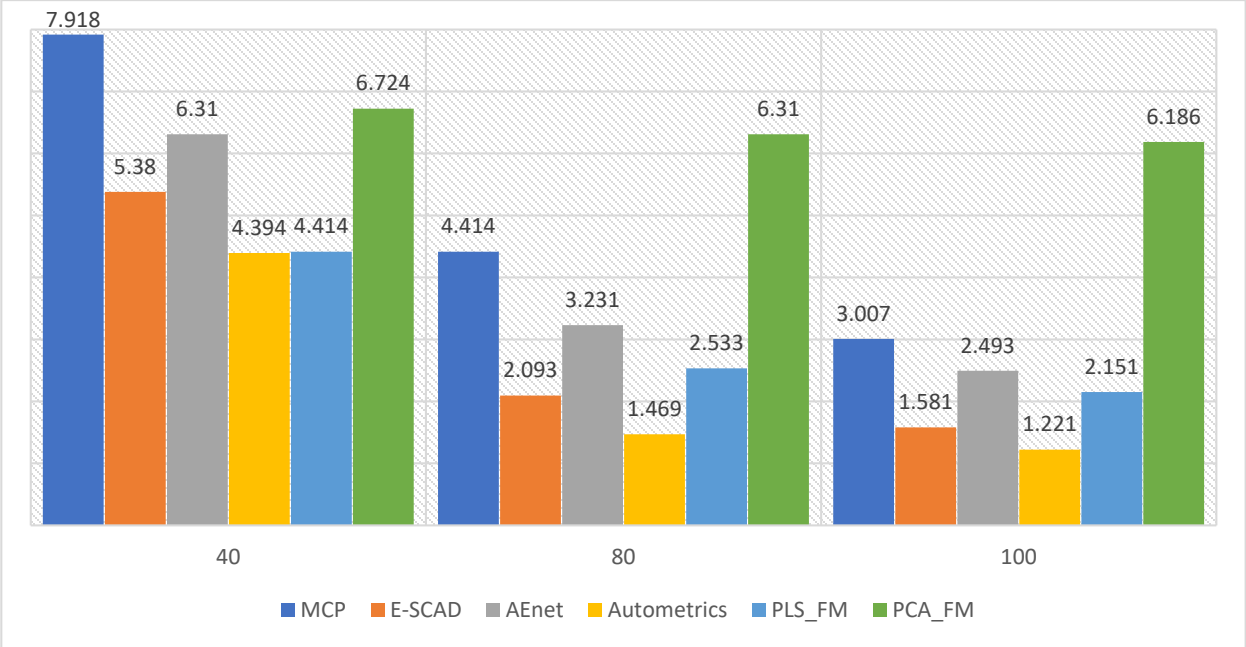
Table 5.4. Forecast comparison under Multicollinearity from Monte Carlo Simulation

Models	$\sum$ ***= 0.25, P = 130***		$\sum$ ***= 0.25, P = 150***	
n = 40/80/100	**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
AEnet	5.049/2.628/2.106	4.113/2.123/1.698	6.414/3.303/2.559	5.267/2.669/2.057
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 = 40/80/100	$\sum$ ***= 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
AEnet	6.310/3.231/2.493	5.163/2.615/2.002	5.129/2.524/2.038	4.204/2.029/1.645
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 = 40/80/100	$\sum$ ***= 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
AEnet	3.342/2.425/2.233	2.731/1.957/1.798	2.791/2.090/1.938	2.888/1.682/1.563
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

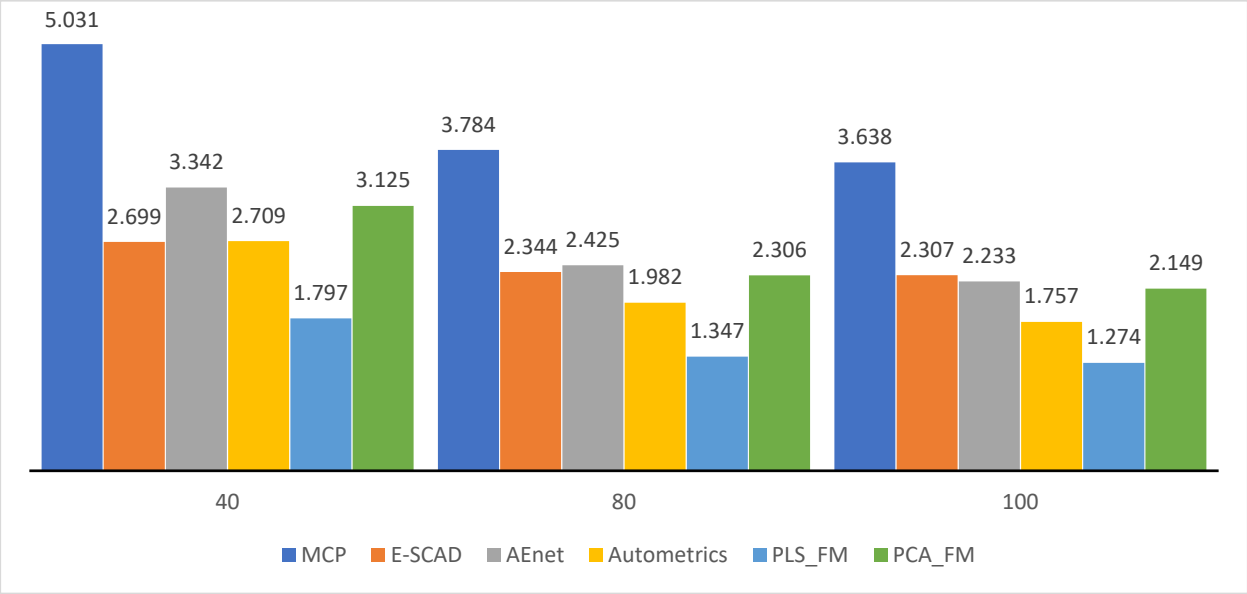
Noted: Bold values indicate a better forecast.


6.86
5.41
2.243
5.8
2
1.355
5.049
2.628
2.106
4.192
1.312
1.189
4.53
3.213
2.727
6.475
5.781
5.695
MCP
E-SCAD
AEnet
Autometrics
PLS_FM
PCA_FM
40
80
100

(a)

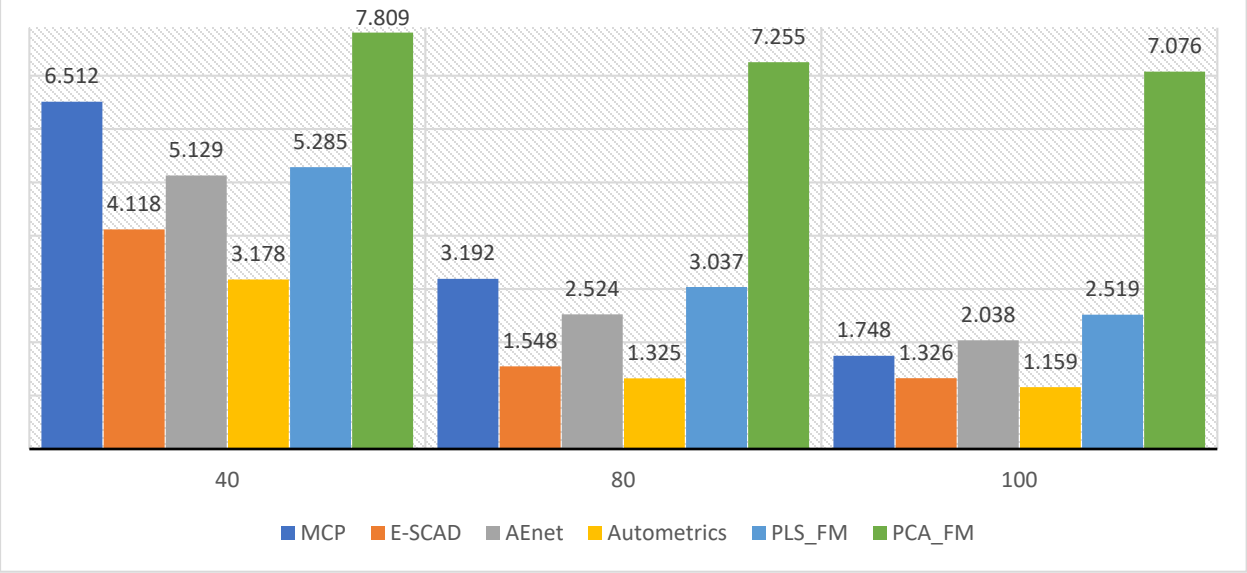

7.918
5.38
6.31
4.394
4.414
6.724
4.414
2.093
3.231
1.469
2.533
6.31
3.007
1.581
2.493
1.221
2.151
6.186
40
80
100
MCP
E-SCAD
AEnet
Autometrics
PLS_FM
PCA_FM

(b)

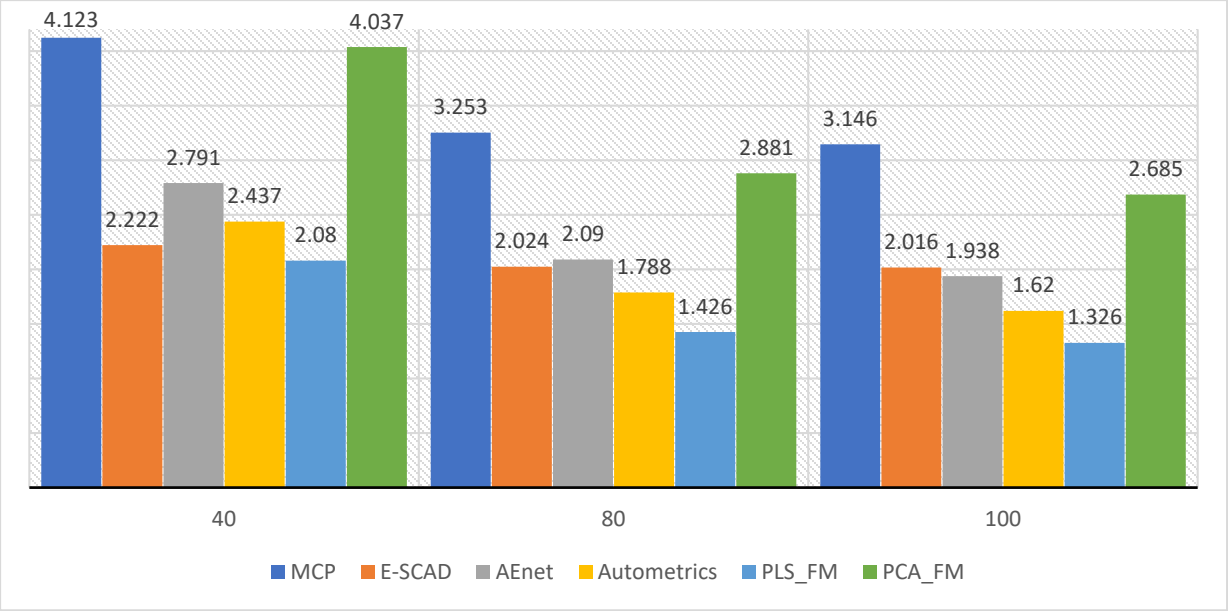


(c)

Figure 5.5: Out of sample root mean squares error across sample sizes, where forecasts are computed from various models when rho = 0.25, rho = 0.5, rho = 0.90 and P = 130.



(a)



(b)

Figure 5.6: Out of sample root mean squares error across sample sizes, where forecasts are computed from several models when rho = 0.9, P = 130(a) and P =150(b).

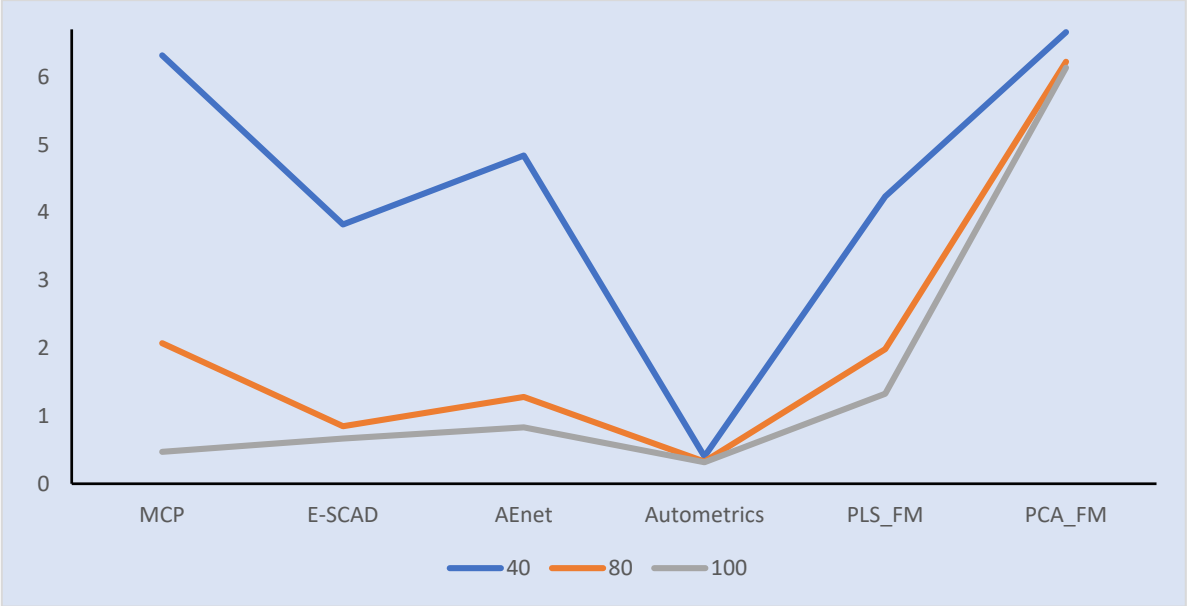
5.2.1.2. SCENARIO-II

Based on RMSE and MAE, the forecasting capabilities of Autometrics is superior to all its competitor counterparts in the presence of heteroscedasticity. In contrast, the MCP and E-SCAD perform poorly using a small sample size, but as we expand the data window (large sample size), their forecasting performance dramatically improves. It indicates that penalized regression models require a large number of data points to provide accurate forecasts. Evidence was found from Fig. 5.7(a, b) and 5.8(a, b) that the forecasting accuracy of all the underlying tools was enhanced with increasing data.

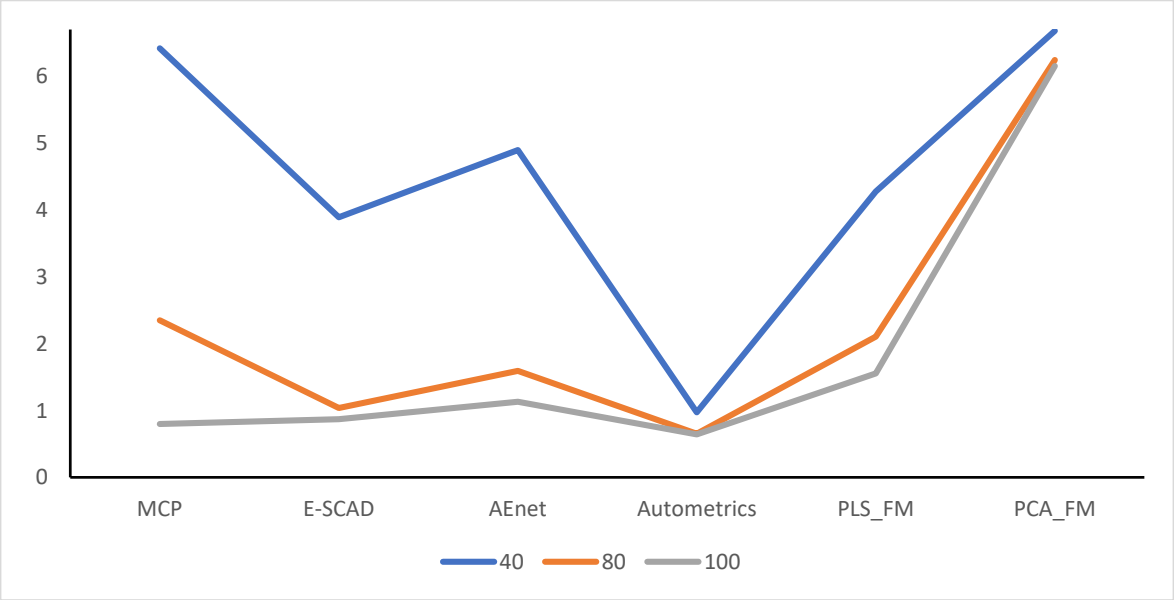
Table 5.5. Forecast comparison under Heteroscedasticity from Monte Carlo Simulation

Models	**π_1 = 0.1/0.3, P = 130**		**π_1 = 0.1/0.3, P = 150**	
n=40/80/100	**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
AEnet	4.840/1.280/0.832	3.961/1.033/0.669	6.106/2.059/1.203	4.993/1.668/0.968
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=40/80/100	**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
AEnet	4.897/1.593/1.132	4.009/1.284/0.911	6.144/2.334/1.514	5.024/1.891/1.218
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=40/80/100	**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
AEnet	4.989/1.980/1.509	4.087/1.594/1.215	6.208/2.683/1.916	5.078/2.172/1.541
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 indicate a better forecast.

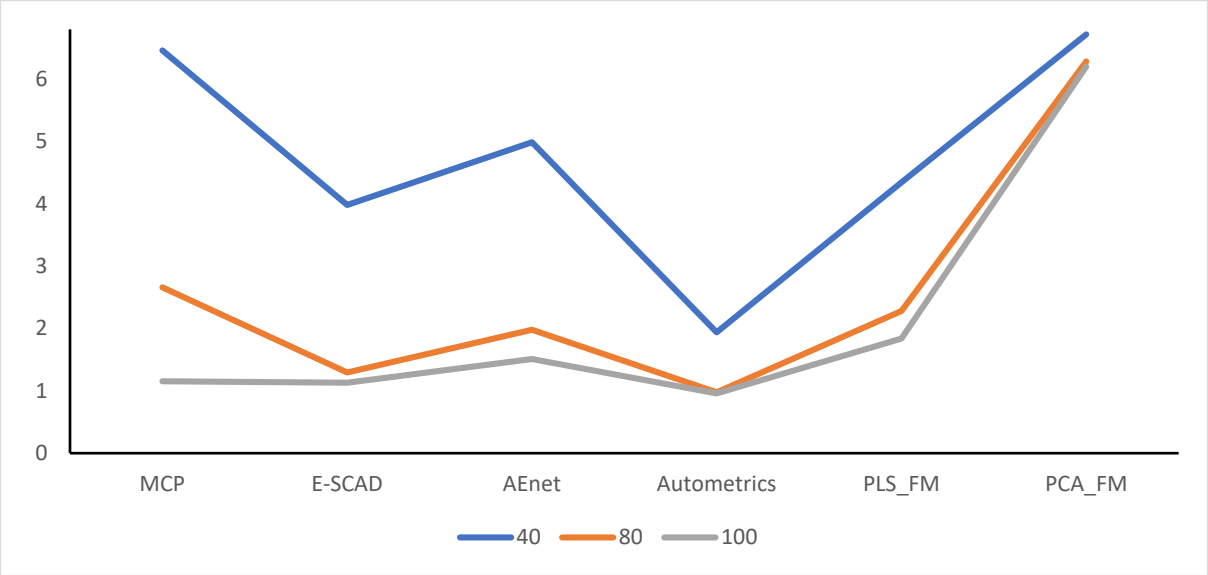


(a)

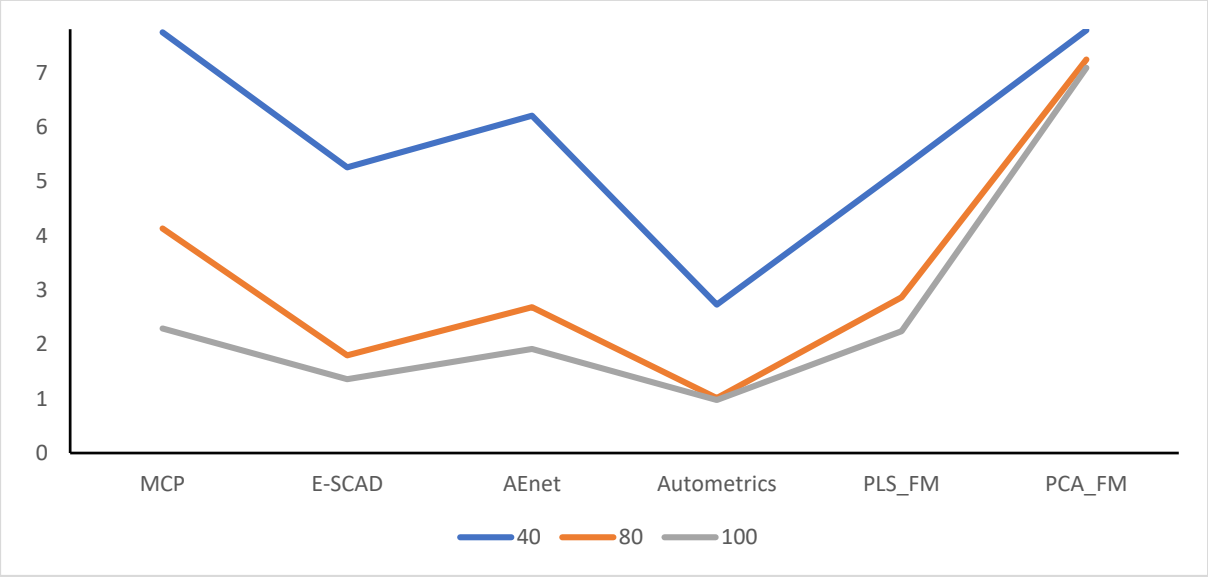


(b)

Figure 5.7: Out of sample root mean squares error across sample sizes, where forecasts are computed from several models when π_1 = 0.1/0.3(a), π_2 = 0.2/0.6(b) and P =130.



(a)



(b)

Figure 5.8: Out of sample root mean squares error across sample sizes, where forecasts are computed from several models when π_1 = 0.3/0.9, P =130(a) and P =130(b).

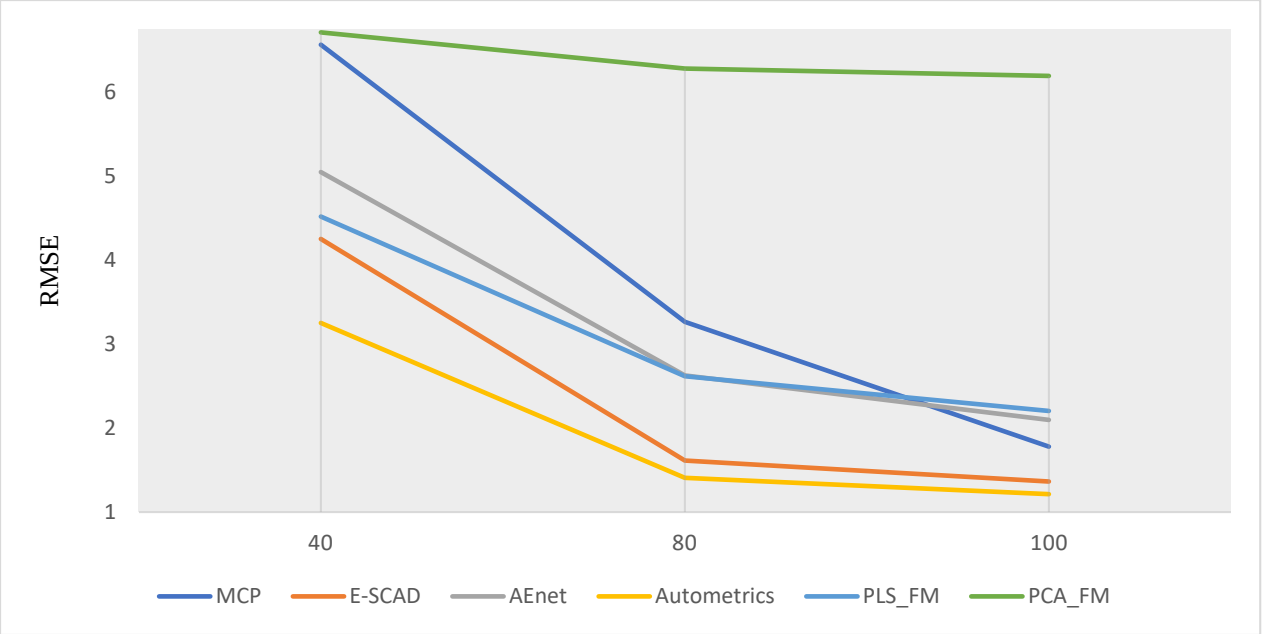
5.2.1.3. SCENARIO-III

Across low and moderate autocorrelation, the Autometrics showed outstanding forecasting performance despite increasing the number of candidate variables (relevant and irrelevant). The E-SCAD remains a good competitor, particularly in case of more observations. While considering the extreme Autocorrelation, E-SCAD provided the lowest RMSE and MAE as compared to its competitor counterparts, regardless of the number of predictors to be used (130/150). Autometrics is still a good competitor. The forecasting performance of all methods improves with augmenting the sample size, which is observed under different schemes of Autocorrelation, shown in Fig. 5.9(a, b, and c). Similarly, it is also found that growing the size of Autocorrelation has an adverse influence on forecasting accuracy, as noted in Fig. 5.10.

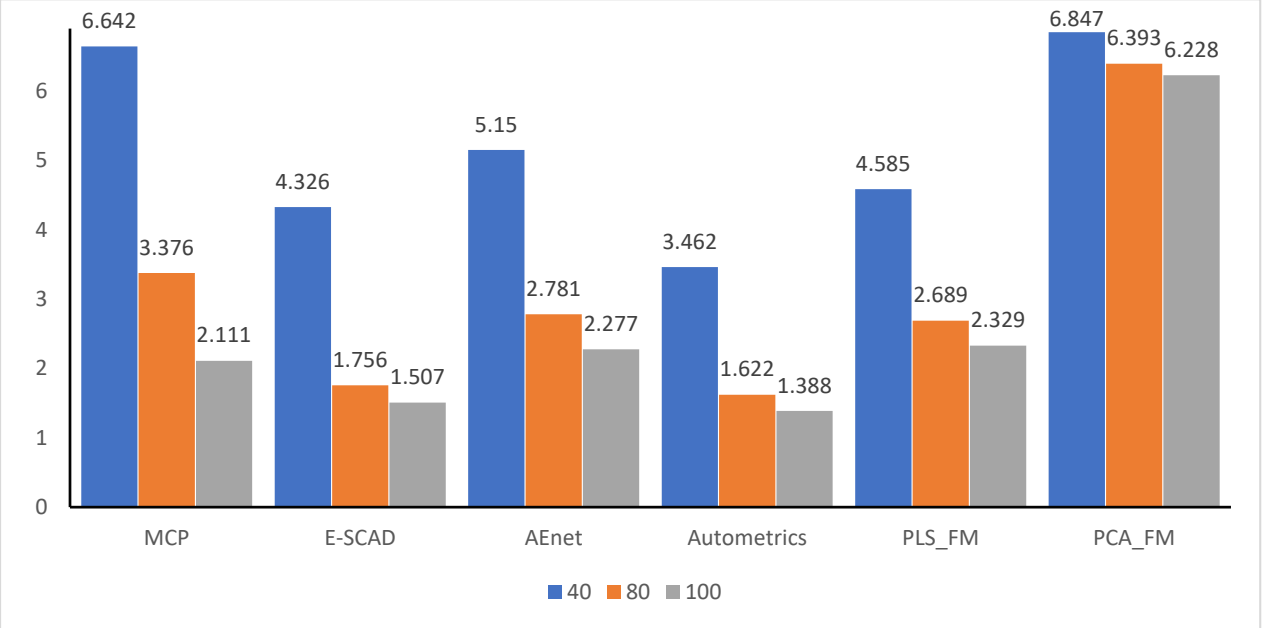
Table 5.6. Forecast comparison under Autocorrelation from Monte Carlo Simulation

Models	***ρ* = 0.25, *P* = 130**		***ρ* = 0.25, *P* = 150**	
n=40/80/100	**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
AEnet	5.049/2.628/2.098	4.113/2.123/1.693	6.213/3.285/2.583	5.090/2.652/2.078
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=40/80/100	***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
AEnet	5.150/2.781/2.277	4.211/2.250/1.838	6.406/3.475/2.765	5.249/2.809/2.233
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=40/80/100	***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**
AEnet	5.782/4.072/3.796	4.737/3.323/3.095	6.925/4.723/4.266	5.685/3.838/3.472
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

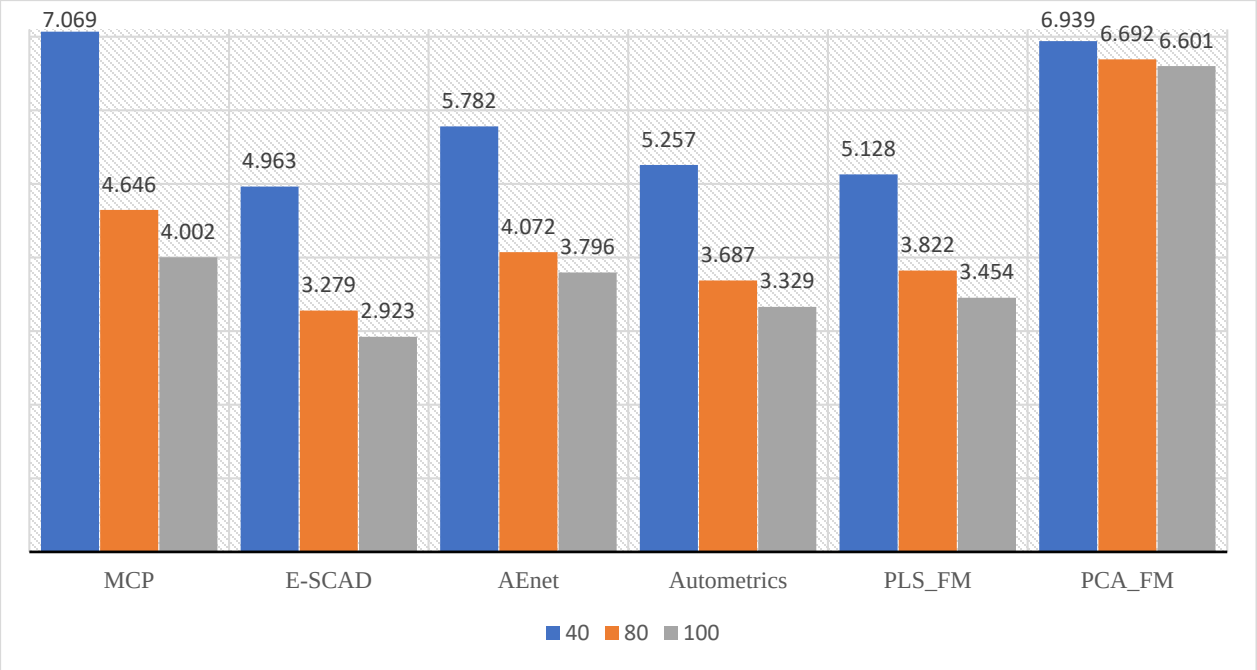
Noted: Bold values indicate a better forecast.


RMSE
6
5
4
3
2
1
40
80
100
MCP
E-SCAD
AEnet
Autometrics
PLS_FM
PCA_FM

(a)


6.642
3.376
2.111
4.326
1.756
1.507
5.15
2.781
2.277
3.462
1.622
1.388
4.585
2.689
2.329
6.847
6.393
6.228
6
5
4
3
2
1
0
MCP
E-SCAD
AEnet
Autometrics
PLS_FM
PCA_FM
40
80
100

(b)



(c)

Figure 5.9: Out of sample root mean squares error across sample sizes, where forecasts are computed from several models when ρ = 0.25(a), ρ = 0.50(b), ρ = 0.90(c) and P =130.

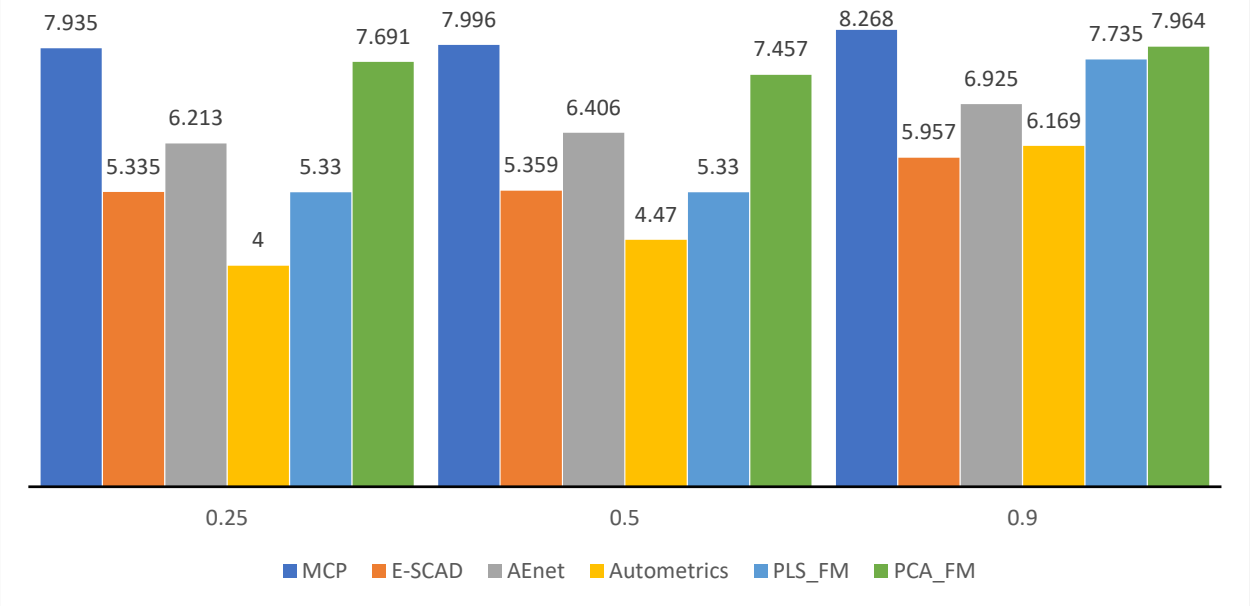


Figure 5.10: Out of sample root mean squares error across the levels of Autocorrelation, where forecasts are computed from several models when n = 40 and P =150.

Chapter 6

Real Data Implications

6.1. Comparison of Variable Selection Methods using Huge Big Data

Using both huge and fat big data, we investigated and compared advanced statistical and machine learning techniques in simulation exercises. Our primary concern is to evaluate the robustness of feature selection and forecasting techniques utilizing a variety of Data Generating Processes (DGPs) and real datasets.

In the discipline of economics, empirical analyses are quite important. The reason for this is that theory without measurement can lead to an inadequate assessment of actual economic challenges. Measurement without economic theory, on the other hand, is inadequate for providing a satisfactory description of how economic forces interact with one another. Neither "theory" nor "measurement" are sufficient to evaluate economics, that is, understanding the relationship between economic variables. Macro-econometric modelling is an established and separate field within the science of economics. It shows a nice composition of economic theories and econometric tools and has been given the most credit for its importance because it gives well-organized frameworks for making policy decisions and planning the economy as a whole.

In this section, we perform some real data analysis in order to support the simulation experiments, which are carried out in the preceding sections. This section consists of four subsections. The first two subsections analyze the variable selection procedures; the last two subsections evaluate the predictive power of the proposed factor model against existing tools using the macroeconomic and financial datasets of Pakistan. The first dataset includes workers' remittance inflow and all its possible determinants.

The question arises, what are the possible determinants of workers' remittance inflow and how do we explore them? There are so many factors that affect the worker's remittances inflow, including economic, financial, political, social, etc. We use two approaches to investigate the factors influencing remittance inflows: literature and economic theories. Some covariates are suggested to be added to the model by economic theories, and a long list of variables has been suggested by studies in the past.

This study includes all the possible determinants based on economic theories and literature to make a general model. In the econometrics literature, such a model is known as the general unrestricted model (GUM).

6.1.1. Data Source

This study collects the yearly data for Pakistan from 1972 to 2020 using different sources such as World Development Indicators (WDI), International Financial Statistics (IFS), International Country Risk Guide (ICRG), and State Bank of Pakistan (SBP). The few missing observations in the data set are replaced by averaging the neighboring observations. Most variables are transformed into logarithmic form to ensure normality. Details regarding the variables have been given in Appendix Table A1. Table A1 describes the variables, symbols, definitions of each variable, and data source.

6.1.2. Correlation matrix

In Fig. 6.1, blue and red colours exhibit positive and negative correlations between the variables. The colours, severity and area of the circles indicate a high pairwise correlation. Besides the right side of the correlogram, the legend colour shows the pairwise correlation. We can see a lot of large blue and red colour circles, which is a sign of a high pairwise correlation.

Figure 6.1 shows that there is high multicollinearity among the predictors using the data period spanning from 1972 to 2020. We noted that in Monte Carlo simulations with high multicollinearity, the AEnet outperformed its rival counterparts in terms of potency and gauge, mainly when the sample size is small. It reveals that AEnet is more robust in such circumstances, and thus we should proceed with AEnet output.

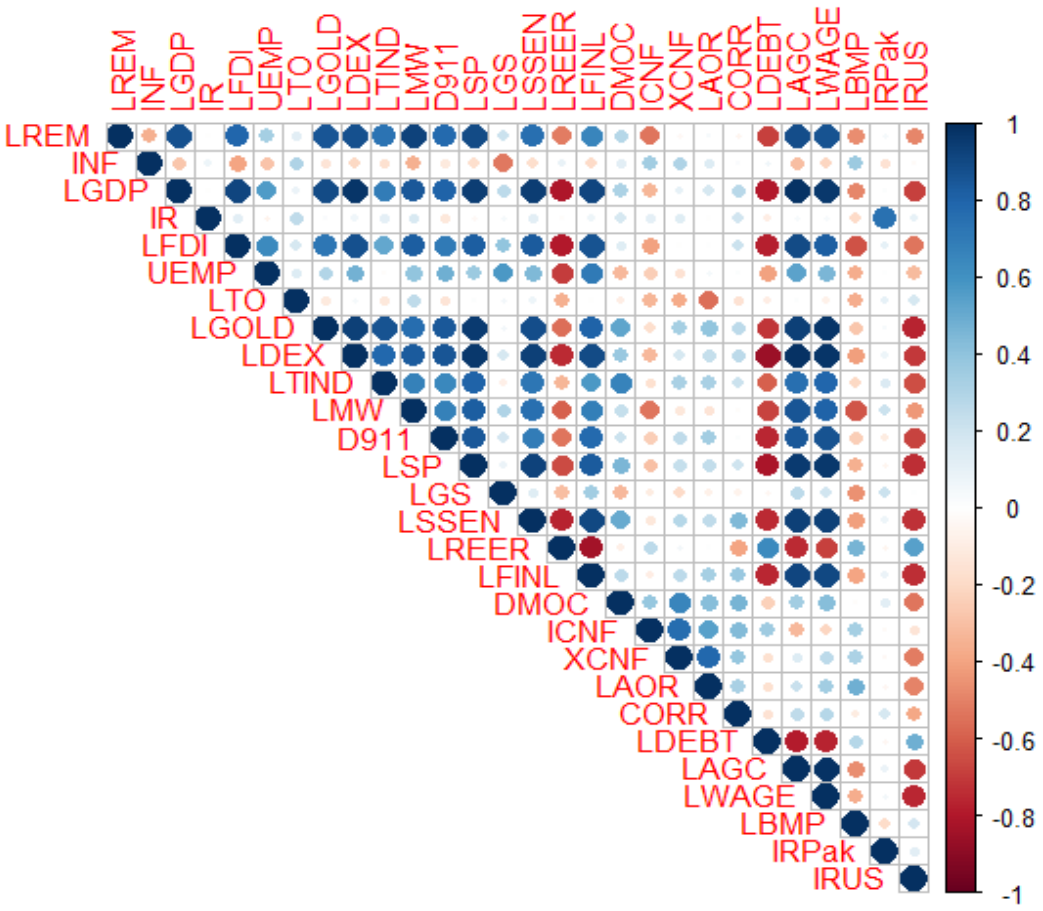


Figure 6.1: Correlation structure among covariates

Table 6.1. Features Selection based on Real Data (Huge Big data)

Variables	MCP	E-SCAD	AEnet	Autometrics
GDP	✓	✓	✓	✓
INF	✓	✓	✓	×
IR	✓	✓	×	×
FDI	✓	×	×	×
UEMP	✓	✓	×	×
TO	✓	✓	✓	✓
GOLD	✓	✓	×	✓
DEX	×	×	×	×
D911	✓	✓	✓	✓
TIND	✓	✓	×	×
MW	✓	✓	✓	✓
SP	✓	✓	×	×
SSEN	✓	×	×	✓
REER	✓	✓	✓	✓
FINL	✓	✓	✓	✓
DMOC	✓	✓	×	✓
ICNF	×	✓	✓	×
XCNF	×	×	×	×
AOR	✓	✓	✓	×
CORR	✓	✓	✓	×
DEPT	✓	✓	✓	✓
GS	✓	×	×	×
IRUS	✓	✓	✓	×
IRPAK	✓	✓	✓	✓
AGC	×	✓	×	×
WAGE	×	×	×	×
BMP	✓	×	×	✓

Table 6.1: Tick marks show the selected variable, and cross marks show the non-selected variable.

Table 6.1 depicts the feature selection based on real data using classical and shrinkage methods. In Table 6.1, the AEnet suggests almost 13 important determinants of workers' remittances among 27 determinants. In contrast, MCP and E-SCAD recommend many unrelated determinants for workers' remittance. In other words, we can conclude that they have over-specified the model, and therefore such models often provide poor forecasts in practice. In a similar way, Autometrics keeps the least number of irrelevant variables, in contrast to MCP and ESCAD, but misses an important variable. However, the right set of covariates can improve forecasting, leading to a low forecast error. Consequently, an accurate forecast can help the government and other sectors in their decision-making.

Referring to simulation results; under the severe case of multicollinearity in simulation, we observed that MCP and ESCAD over-specified the model, whereas the Autometrics approach suffered from under-specification. In contrast to these findings, the results produced by the AEnet approach are very close to the true DGP. When analyzing real data, MCP and ESCAD retained more irrelevant variables, whereas Autometrics dropped the important variable(s). The performance of AEnet showed the same behaviour as shown in the simulation exercise. As a whole, the results show that the empirical application strongly backs up the results of the simulation exercise.

Using another dataset, we use the same procedures for feature selection and compare their performance, as shown in Table 6.2. In each column, we mention the variable name, which is selected by the specific method. In practice, E-SCAD selected 11 predictors, MCP kept one variable, DEX, AEnet selected four variables, and Autometrics kept four predictors. In presence of severe multicollinearity, Enet has shown outstanding performance in simulation exercises.

Thus, the variables selected by Enet are the most useful drivers for the stock market. As we can notice, E-SCAD retained many irrelevant variables that are probably not important for the stock market. MCP, in contrast, under-specifies the true model of the stock market, and such a model often produces biased results.

Table 6.2. Features Selection based on Real Data (Huge Big data)

E-SCAD	**MCP**	**AEnet**	**Autometrics**
GDP	-	GDP	GDP
-	-	-	GS
Debt	-	Debt	Debt
OP	-	-	OP
GFCF	-	-	GFCF
FDI	-	-	-
Gold	-	Gold	-
REER	-	REER	-
IR	-	-	-
DEX	DEX	-	-
TIND	-	-	-
REM	-	-	-

Noted: Selection of important features for Stock market prices.

6.2. Comparison of Variable Selection Methods using Fat Big Data

We analyze the macroeconomic time series data set for Pakistan. The data set consists of 79 aggregated and disaggregated variables collected at a monthly frequency for the period starting 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 is taken from the state bank of Pakistan.

The forecasting model is basically constructed for inflation (INF), where a long list of variables is incorporated as explanatory variables in the model. All the variables (response variables and explanatory variables) are transformed in order to make them stationary before an empirical analysis. Generally, the logarithmic transformation is performed for all non-negative time series that are not already in rate (Stock and Watson, 2012). A complete list of variables is given in the Appendix. Details on the variables used for analysis are given in Appendix Table A2.

Before modelling, we check the multicollinearity among the explanatory variables through a correlation matrix, which is not possible to show here. The correlation matrix showed a strong pairwise correlation in the set of covariates. Moreover, the frequency of the data set under consideration is monthly, so there is a huge likelihood of autocorrelation in the data.

The real data based comparison is given in Table 6.3. It can be seen from the table that E-SCAD selects 17 features, the MCP retains 8 features, the AEnet retains 4 features, and the Autometrics holds 5 features out of the entire set of 78 features. To relate these results with the output of simulation experiments, the E-SCAD has higher potential in contrast to competing approaches while retaining the relevant features in the presence of high pairwise correlation in the predictor set (severe multicollinearity). It was also shown through simulation experiments that regardless of such good performance, the E-SCAD keeps irrelevant variables as well. Analyzing the real dataset, it is also noticed that out of 17 features, some are more likely to be irrelevant. However, the rival approaches have dropped the important features (as observed in simulation), which in turn leads to biased estimates. In view of all this, the output of the real data analysis supports the simulation results.

Table 6.3. Features Selection based on Real Data (Fat Big data)

E-SCAD	MCP	AEnet	Autometrics
Wholesale Price Index	Wholesale Price Index	Wholesale Price Index	Wholesale Price Index
Sensitive Price Index	-	Sensitive Price Index	Sensitive Price Index
Federal Government Indirect Tax (Excise Tax)	Federal Government Indirect Tax (Excise Tax)	-	-
Federal Government Indirect Tax (Customs)	-	-	-
Old Foreign Currency Accounts	Old Foreign Currency Accounts	-	-
Call Money Rate	-	-	-
Nominal effective exchange rate	Nominal effective exchange rate	-	Nominal effective exchange rate
Real effective exchange rate	Real effective exchange rate	Real effective exchange rate	Real effective exchange rate
Japanese Yen (Monthly Average)	Japanese Yen (Monthly Average)	-	-
Other Deposits with State Bank of Pakistan	Other Deposits with State Bank of Pakistan	Other Deposits with State Bank of Pakistan	Other Deposit with State Bank of Pakistan
Currency in Tills of Scheduled Banks	-	-	-
Time Deposits	-	-	-
Caustic Soda	-	-	-
Chlorine Gas	-	-	-
Lime Stone	Lime Stone	-	-
Crude Oil	-	-	-
Natural Gas	-	-	-

6.3. Out-of-Sample Forecasting Comparison using Huge Big Data

For real data analysis, we focus on two datasets: macroeconomic data and financial market data. The macroeconomic and financial dataset includes workers remittances inflow (which is already discussed in section 6.1) and stock market prices. Stock market prices are determined by many factors, including economic, financial, political, social, etc. To make a GUM, this study looks at all the possible determinants based on theories and literature.

6.3.1. Data Source

This study collects the annual data for Pakistan from 1972 to 2020. The data is sourced from the World Development Indicators (WDI), international financial statistics (IFS), international country risk guide, and the state bank of Pakistan. The few missing observations in the data set are replaced by averaging the neighboring observations.

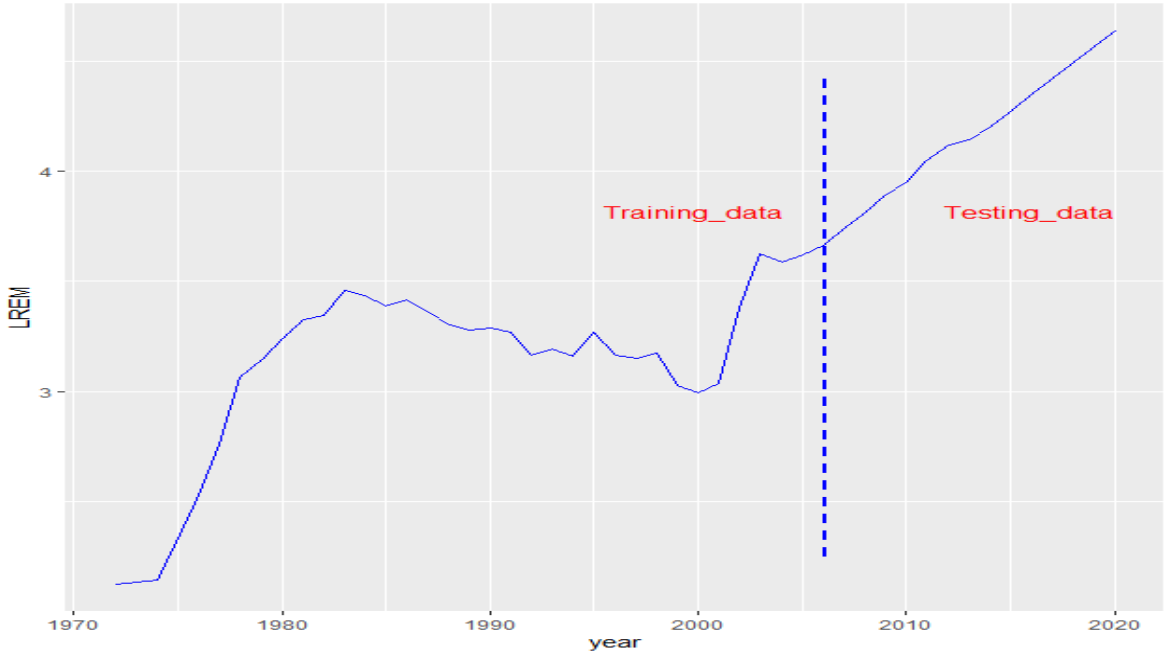


Figure 6.2: Trend of Workers' Remittance series (in the log) against time.

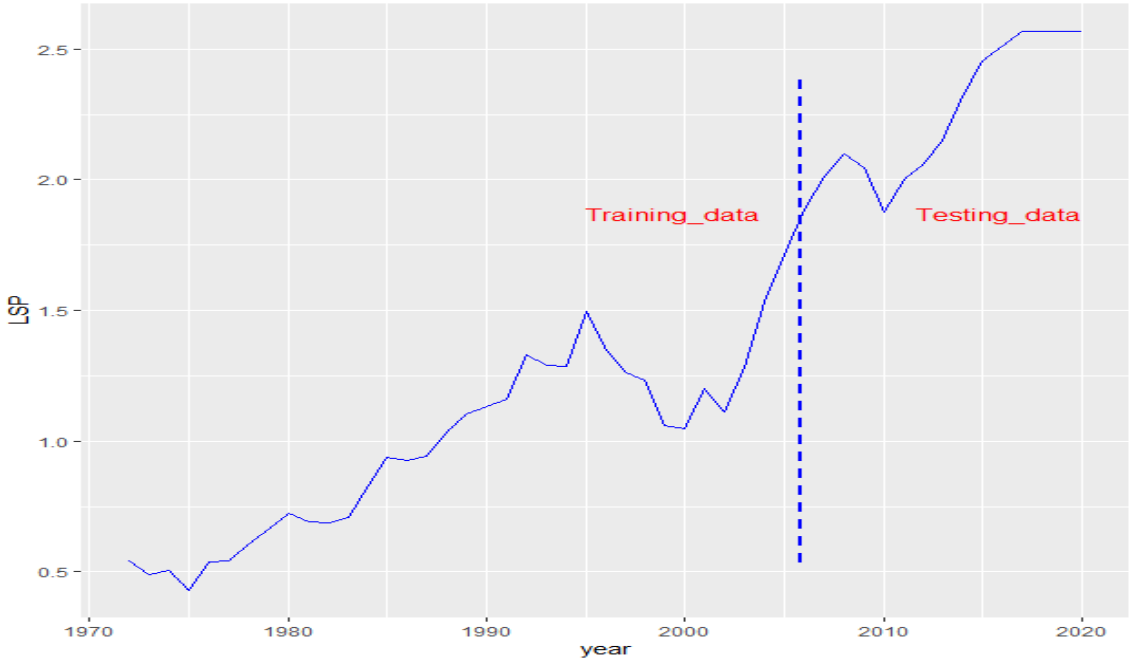


***Figure 6.3**:* Trend of Stock Prices series (in the log) against time.

6.3.2. Correlation Matrix

For empirical analysis, we split the data set into parts: observations from 1972–2007 are utilized to train the models, and the remaining data (testing data) is used to evaluate their forecasting performance, which is provided in Figures 6.2 and 6.3. But before going to compute the forecast error, we discover the correlation structure among covariates through the visualization approach**.** The plot of pairwise correlation for the workers' remittances dataset is provided in Fig. 6.4, where **blue and red colours exhibit positive** and negative correlations, respectively. The colour severity and the area of the circle are directly associated with **correlation coefficients**. On the right side of the **correlogram**, the legend color shows the **correlation coefficients** and the corresponding colours. We can observe that there are many dark colour circles in blue and red, which clearly illustrates the high pairwise correlation. In other words, we can conclude that there is high multicollinearity among the predictors of a dataset. Likewise, the pairwise correlation of the second dataset related to stock market determinants is provided in Table 6.4. Here, the statistical

significance of the correlation is indicated by stars. It can be seen from the table that many pairs of variables are highly correlated, which ensures the case of severe multicollinearity. Figure 6.5 reveals that the distribution of stock market data is almost symmetric. As we have noted in simulation experiments, in the presence of high multicollinearity (what we observed in the real datasets), the PLS-based factor model outperformed the other procedures in terms of producing a low forecast error for n = 80. It reveals that a PLS based factor approach is more robust in such circumstances.

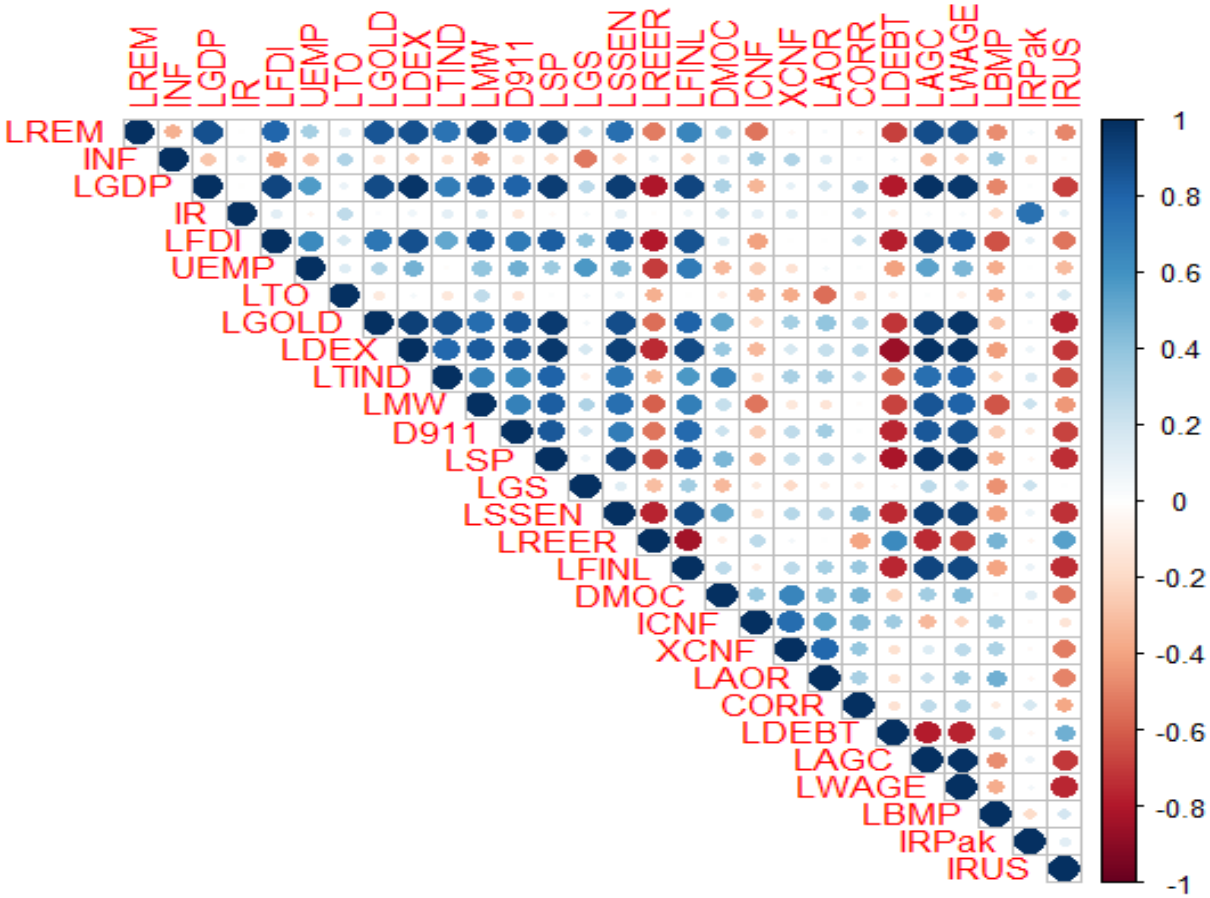


Figure 6.4: Pairwise correlation using the determinants of workers' remittance

Table 6.4: Pairwise Correlation using the Determinants of Stock Market

	LGDP	*LGS*	*LDEBT*	*OP*	*GFCF*	*LFDI*	*UEMP*	*LTO*	*LGOLD*	*LREER*	*LFINL*	*INF*	*IR*	*LDEX*	*LTIND*	*LREM*	*MS*
LGDP																	
LGS	0.263																
LDEBT	0.795***	-0.046															
OP	0.561***	0.028	0.667***														
GFCF	-0.176	0.181	0.155	-0.393**													
LFDI	0.919***	0.393**	0.773***	0.507***	0.076												
UEMP	0.568***	0.571***	-0.406**	0.331*	0.024	0.634***											
LTO	0.089	0.053	-0.094	-0.055	0.505***	0.174	0.143										
LGOLD	0.899***	0.052	0.715***	0.670***	-0.441**	0.726***	0.294*	-0.116									
LREER	0.809***	-0.309*	0.636***	-0.342*	0.008	0.799***	0.690***	0.354*	0.554***								
LFINL	0.918***	0.348*	0.758***	0.564***	-0.262	0.863***	0.701***	-0.007	0.807***	0.833***							
INF	-0.280	0.522***	0.063	0.080	-0.168	-0.397**	-0.284*	0.301*	-0.150	0.096	-0.193						
IR	-0.007	0.044	-0.092	0.179	0.017	0.128	-0.062	0.256	0.024	-0.011	0.076	0.070					
LDEX	0.974***	0.171	0.856***	0.667***	-0.265	0.880***	0.475***	0.044	0.938***	0.740***	0.899***	-0.209	0.069				
LTIND	0.691***	-0.081	0.597***	0.483***	0.457***	0.515***	-0.014	-0.120	0.866***	-0.337*	0.574***	-0.165	0.117	0.781***			
LREM	0.873***	0.216	0.685***	0.575***	-0.105	0.792***	0.336*	0.124	0.860***	0.519***	0.658***	0.350*	0.007	0.874***	0.735***		
MS	-0.103	0.073	0.056	0.020	0.384**	0.068	0.178	0.080	-0.160	0.061	-0.037	-0.146	0.062	-0.119	-0.188	-0.101	

Computed correlation used pearson-method with listwise-deletion.

Noted: Statistical significance of the pairwise correlation is represented by stars, as 1, 5, and 10 percent are indicated by (***), (**), and (*).

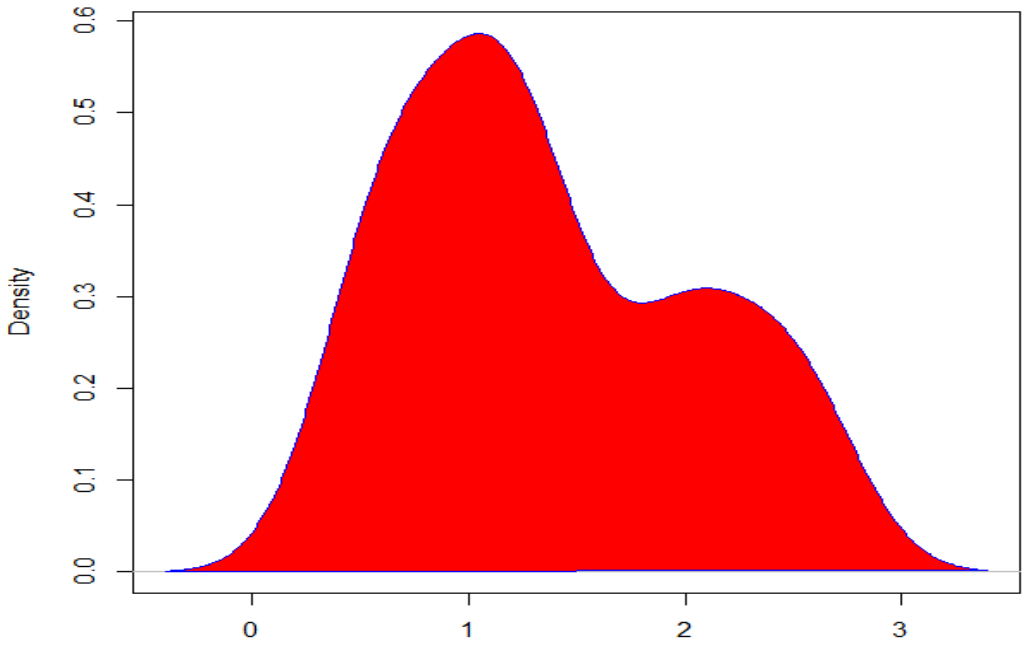
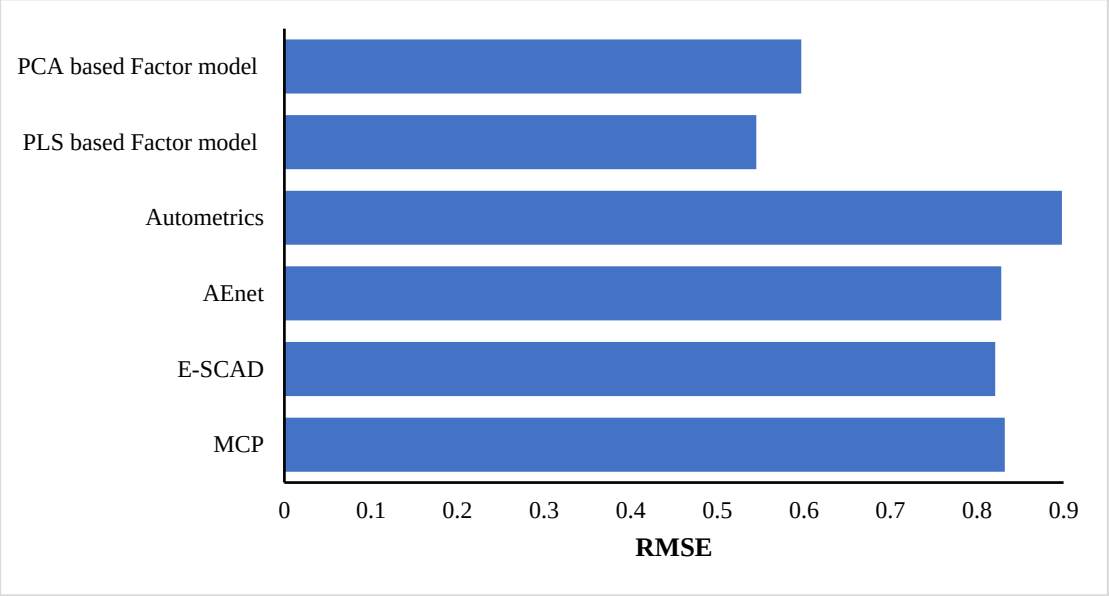


Figure 6.5: Density Plot of Pakistan Stock Prices.

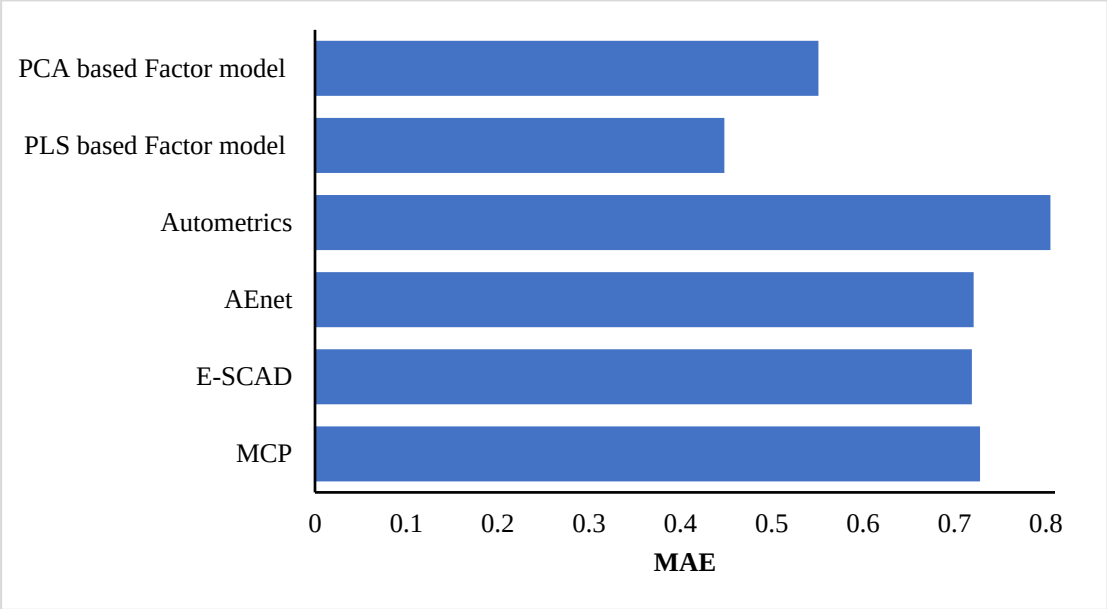
6.3.3. Forecast Comparison Based on Dual Real Datasets

Figures (6.6 and 6.7) present the forecasting experiment across different forecasting procedures for one of the core macroeconomic variables (inflation), and the second core variable is a financial variable of interest (stock market prices). The forecasting accuracy is given as the RMSE and MAE, which are represented in our case by a bar chart against different methods. The smaller the length of a bar, the better the forecast attained by a model, comparatively. The length of a bar in Fig. 6.6(a, b) indicates that the PLS-based factor model outperformed the competing methods in the out-of-sample forecast. It illustrates that the PLS-based factor model (proposed model) has better predictive power than other competitor models, in terms of having the lowest forecast errors in multi-step-ahead forecasts (2008 to 2020). Seeing another figure 6.7(a, b), we obtained a similar outcome. The proposed model shows an outstanding prediction against the competing approaches. It is noted that Autometrics showed the worst performance, as it showed in the simulation under

similar circumstances (extreme multicollinearity). The real data results support the simulation results under both real datasets.

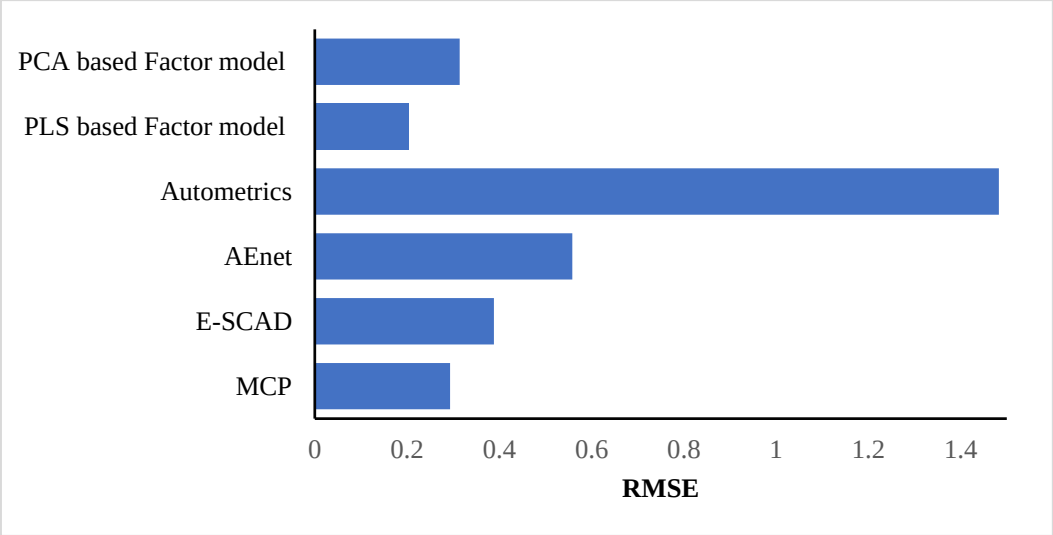


(a)

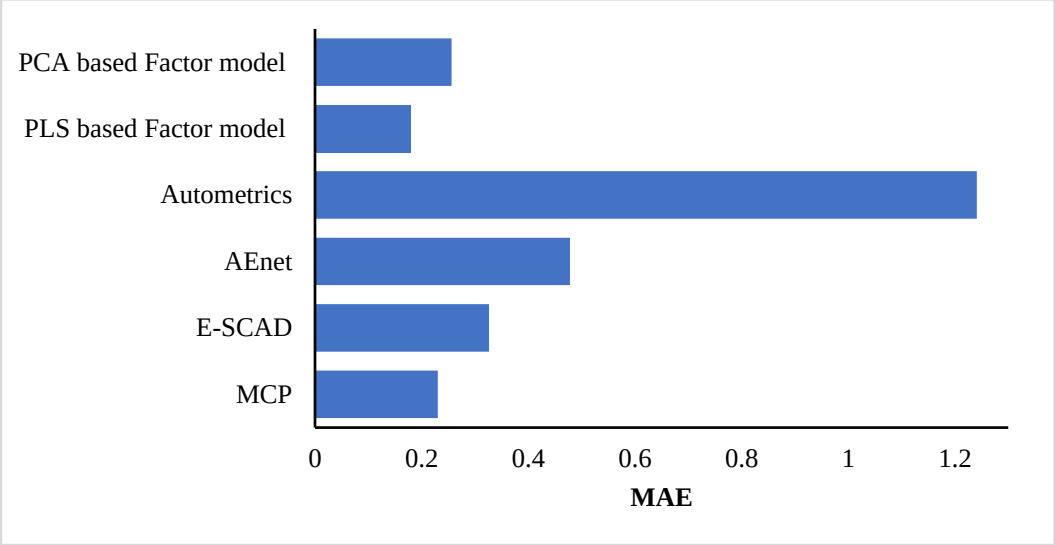


(b)

Figure 6.6: The proposed model versus the baseline models (Workers' Remittance series)



(a)



(b)

***Figure 6.7*: The proposed model versus the baseline models (Stock prices series)**

6.4. Out-of-Sample Forecasting Comparison using Fat Big Data

We analyze the macroeconomic time series data set for Pakistan. This dataset has already been used in section 6.2, where a detailed explanation has been given regarding the variables, source, and frequency of data. Details on the variables used for analysis are given in Appendix Table B1.

6.4.1. Inflation Forecasting

The dataset is divided into two parts with the intention of facilitating out-of-sample forecast accuracy. For model estimation, we utilize the data from January 2013 to February 2019, and March 2019 to December 2020, for assessing the models' multistep-ahead post-sample prediction accuracy. The inflation time series plotted in Fig. 6.8 is divided by a vertical blue dotted line, where the training part is used for model estimation and the second part (testing data) is used for out-of-sample prediction.

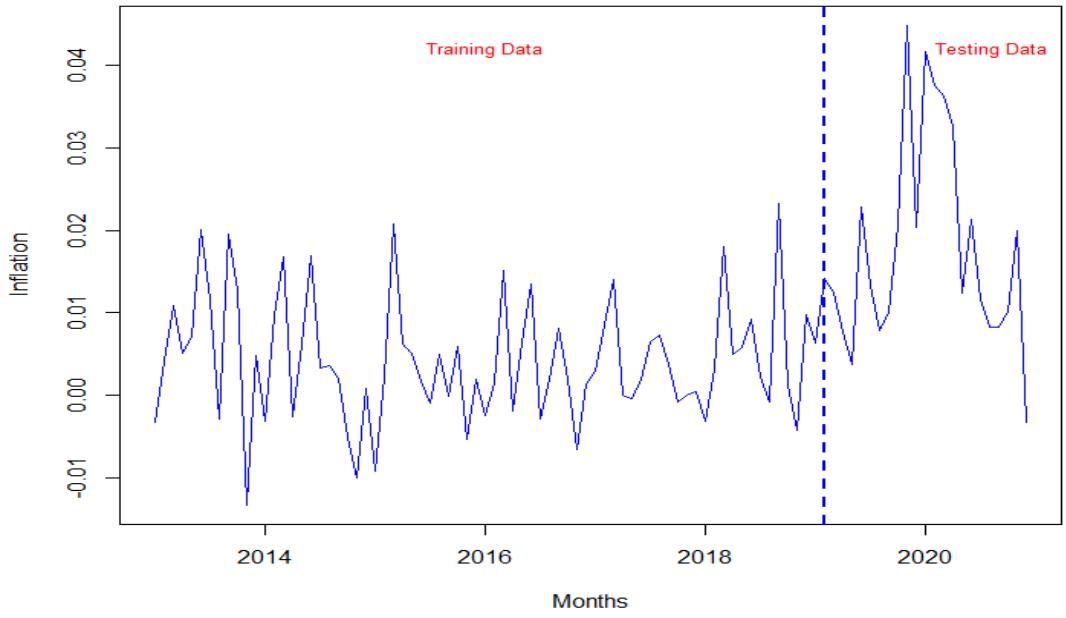
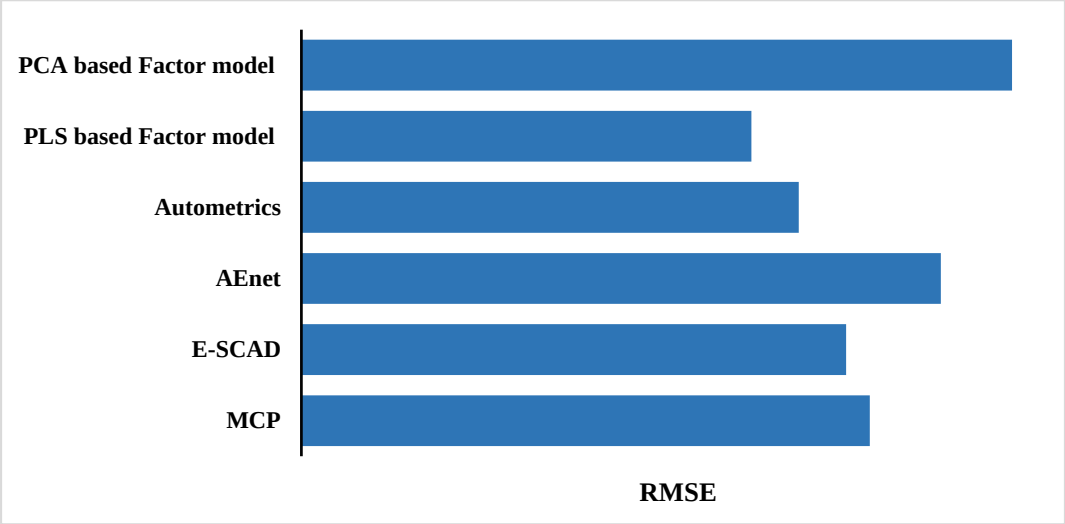


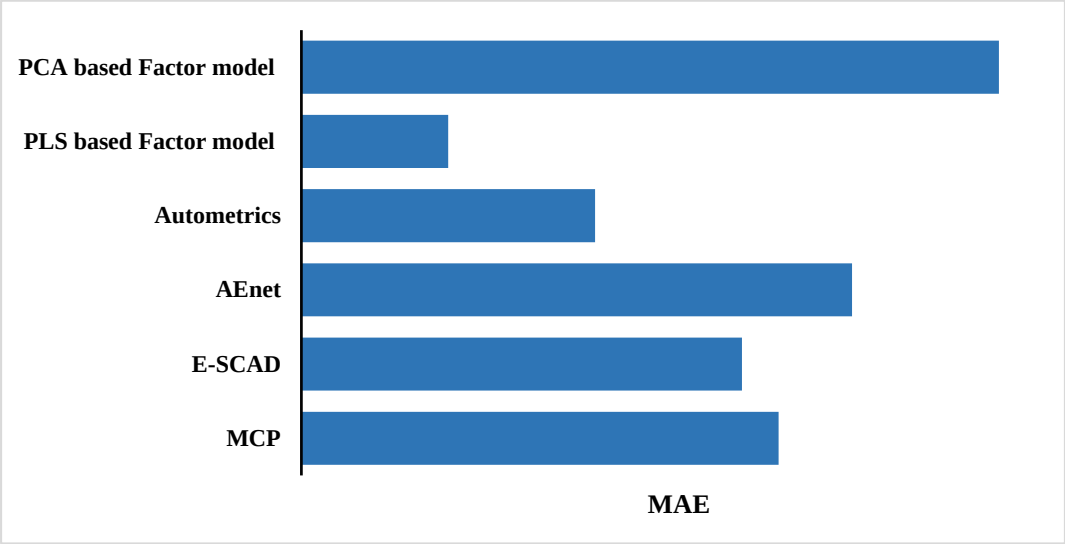
Figure 6.8: Monthly inflation detrended series against time

Fig. 6.9 presents 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, which is represented in our case by a bar chart against different methods. The smaller the length of a bar, the better the forecast attained by a model, comparatively. By observing the length of the bar given in Fig. 6.9, we can infer that the PLS based factor model is superior to

its rival counterparts in post sample forecast. In other words, the forecasted values are close to the observed data on inflation. In contrast, Autometrics produces a good forecast but is not as satisfactory as PLS based factor model.



(a)



(b)

Figure 6.9: The proposed model versus the baseline models (Inflation series)

Chapter 7

Conclusion, Limitations and Future Direction

The primary objectives of statistical learning are to ensure high prediction accuracy and identify relevant predictive variables. Variable selection is crucial when the representation of the true underlying model is sparse. Identifying important predictors will enhance the predictive ability of the fitted model. Literature discusses a variety of methods for selecting variables, but each method selects a distinct subset of variables and performs differently under distinct conditions. Through comparison, we can evaluate their relative performance. The first objective of this study is to compare different variable/model selection tools, namely Autometrics, Adaptive Elastic net, Elastic-Smoothly Clipped Absolute Deviation, and Minimax Concave Penalty using huge big data and Fat Big data. The comparison is made under different scenarios, including Multicollinearity, Heteroscedasticity, and Autocorrelation with varying sample sizes and the candidate set of variables (relevant and irrelevant). The study performed Monte Carlo experiments to compare all methods in terms of variable selection using potency and gauge. First, we discuss the results obtained from huge big data:

Considering the cases of low and moderate multicollinearity as well as low and moderate autocorrelation, all the techniques often retain all the relevant predictor variables. However, in terms of gauge, the MCP and E-SCAD keep many irrelevant predictors, and thereby over-specify the models under the same scenarios. The AEnet retains more than 93 percent of the correct variables in the presence of extreme multicollinearity. However, the potency of the remaining techniques, specifically MCP and E-SCAD, tends towards unity with increasing sample size, capturing the massively irrelevant predictors as well.

Considering the higher level of autocorrelation, E-SCAD has shown good performance in the selection of relevant variables under a small sample, but the same method collapsed under a gauge. On the other hand, Autometrics and AEnet performed better in gauge and frequently held less than 5% of irrelevant variables. In presence of heteroscedasticity, all adopted techniques often hold all relevant variables, but also suffer from over-specification problems, except AEnet and Autometrics which avoid irrelevant predictors and identify the true model precisely.

Secondly, we delineate the findings coming from Fat Big data:

In case of low multicollinearity, the Autometrics often retain the true DGP. By enlarging the covariate window, the E-SCAD frequently outperforms the competitors in terms of potency, though in terms of gauge, the Autometrics performance is the best among the competitors. The moderate level of multicollinearity declines the potency of Autometrics and AEnet, while improving the MCP and E-SCAD. In presence of high multicollinearity, E-SCAD showed an outstanding performance.

In case of low and moderate heteroscedasticity, it can be seen that the potency of Autometrics is usually higher than competitor tools, and it retains fewer irrelevant variables. increasing the strength of heteroscedasticity, which in turn declines the potency of Autometrics in comparison to shrinkage techniques. Whatever the level of Autocorrelation**,** the potency of E-SCAD is often higher than the competitive counterparts. In terms of gauge, the Autometrics showed good performance, which circumvents the inclusion of irrelevant variables. The AEnet is a good competitor to Autometrics in gauge, asymptotically.

The second goal of our study is to compare how good the proposed factor model (PLS based factor model) is at making predictions with existing tools under the same conditions.

In the presence of low and moderate multicollinearity as well as low and moderate autocorrelation, the MCP often produced better forecasts than the rival methods for large sample size. Despite this, E-SCAD frequently outperforms competing methods for small samples. Considering the case of extreme collinearity, the PLS-based factor model is superior in case of a small sample. In case of extreme autocorrelation, the E-SCAD outperformed the rival techniques except at n = 400, where the MCP induced a more accurate forecast. In presence of heteroscedastic errors, MCP remains an effective tool.

Considering the case of low and moderate multicollinearity, the forecasting performance of Autometrics is more promising than its competitive counterparts. In presence of extreme multicollinearity, the factor approach based on PLS outperformed rival counterparts. Similarly, in presence of heteroscedastic errors, the forecasting capability of Autometrics is superior to all competitors. Across the low and moderate sizes of autocorrelation, the predictive power of Autometrics remained higher despite increasing the candidate set of variables. In terms of extreme autocorrelation, the E-SCAD produced the lowest forecast errors.

To achieve the third and last objective of our study by analysing the real datasets of Pakistan.
On the application side, we take the workers' remittance data along with its twenty-seven determinants, spanning from 1972 to 2020. The AEnet keeps 13 predictors, the Autometrics holds 12 predictors, and the MCP and E-SCAD have over-specified the models due to retaining many irrelevant determinants affecting the workers' remittance.

Complementing the simulation exercises, we analyse the macroeconomic time series data set for Pakistan. The data set consists of 79 aggregate and disaggregated variables collected at a monthly frequency for the period starting from 2013 to 2020. The dataset covers the fiscal sector, real

sector, financial sector, monetary sector, and external sector of the economy of Pakistan. The data is taken from the State Bank of Pakistan. The findings conclude that E-SCAD selects 17 features, the MCP retains 8 features, the AEnet retains 4 features, and the Autometrics holds 5 features out of the entire set of 78 features. To relate these results with those of simulation experiments, the E-SCAD has higher potential in contrast to rival counterparts while retaining the relevant features in the presence of autocorrelation and multicollinearity problems.

For empirical applications, macroeconomic and financial datasets are used. To compare the forecasting performance of the included methods, we divide the data into two parts, i.e., data over 1973–2007 as training data and data over 2008–2020 as testing data, using both datasets. All methods are trained on training data and subsequently, their performance is evaluated through testing data. Based on RMSE and MAE, the PLS based factor model showed superior performance against its competitor counterparts.

Using the same dataset (inflation data), we divide the dataset into two parts in order to facilitate out-of-sample forecast accuracy. For model estimation, we utilise the data from January 2013 to February 2018, and March 2019 to December 2020, for assessing the models' multistep-ahead post-sample prediction accuracy. It can be inferred that a PLS based factor model is superior to its rival counterparts in post-sample forecasting.

7.1. Limitations and Future Direction

The few limitations of this study are that it only focuses on linear models. Secondly, because simulation analysis is specific to the underlying setup, it is difficult to generalize the results of the simulation. Thirdly, the simulation part of this study is restricted to Gaussian distributed errors, but in practice, it is not essential that the errors of a model are always normal. Similarly, we have

evaluated the selected models under multicollinearity, autocorrelation, and heteroscedasticity. What if these three problems occur simultaneously? Hence, this research can be expanded to discover the forecasting performance of advanced statistical and machine learning techniques under non-normal residuals as well as missing observations in the data set. Moreover, it is also possible to consider the lagged variables and compare these tools in terms of forecasting and variable selection. This study can be expanded to examine the performance of non-linear and non-parametric algorithms like artificial neural networks, random forests, support vector machines, etc. Finally, researchers should focus on addressing the aforementioned problems simultaneously.

References

Aamir, M., & Ali Shah, S. Z. (2018). Determinants of stock market co-movements between Pakistan and Asian emerging economies. *Journal of Risk and Financial Management*, *11*(3), 32.

Abbas, F., Masood, A., & Sakhawat, A. (2017). What determine remittances to Pakistan? The role of macroeconomic, political and financial factors. *Journal of policy modeling*, *39*(3), 519-531.

Abbas, S. (2020). Impact of oil prices on remittances to Pakistan from GCC countries: evidence from panel asymmetric analysis. *OPEC Energy Review*, *44*(2), 205-223.

Adams, S. (2009). Foreign direct investment, domestic investment, and economic growth in Sub-Saharan Africa. *Journal of policy modeling*, *31*(6), 939-949.

Ahmed, J., & Martinez-Zarzoso, I. (2014). What drives bilateral remittances to Pakistan? A gravity model approach. *A Gravity Model Approach (June 3, 2014)*.

Ahmed, J., & Martínez-Zarzoso, I. (2016). Do transfer costs matter for foreign remittances? A gravity model approach. *Economics*, *10*(1).

Ahmed, J., Mughal, M., & Martínez-Zarzoso, I. (2021). Sending money home: Transaction cost and remittances to developing countries. *The World Economy*, *44*(8), 2433-2459.

Ahumada, H., & Cornejo, M. (2016). Forecasting food prices: The case of corn, soybeans and wheat. *International Journal of Forecasting*, *32*(3), 838-848.

Akçay, S. (2021). Are oil prices and remittance outflows asymmetric? Evidence from Saudi Arabia. *Energy Research Letters*, *2*(1), 18948.

Akçay, S., & Karasoy, A. (2019). The asymmetric impact of oil prices on remittances: evidence from India. *OPEC Energy Review*, *43*(3), 362-382.

Algamal, Z. Y., & Lee, M. H. (2015). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Computers in biology and medicine*, *67*, 136-145.

Ali, S., Khan, H., Shah, I., Butt, M. M., & Suhail, M. (2021). A comparison of some new and old robust ridge regression estimators. *Communications in Statistics-Simulation and Computation*, *50*(8), 2213-2231.

Armah, N. A., & Swanson, N. R. (2010a). Diffusion index models and index proxies: Recent results and new direction. European Journal of Pure and Applied Mathematics, 3, 478–501.

Armah, N.A., Swanson, N.R. (2010b). Seeing inside the black box: using diffusion index methodology to construct factor proxies in large scale macroeconomic time series environments. *Econometric Reviews 29*, 476–510.

Artis, M. J., Banerjee, A., & Marcellino, M. (2005). Factor forecasts for the UK. *Journal of forecasting*, *24*(4), 279-298.

Arun, T., & Ulku, H. (2011). Determinants of remittances: The case of the South Asian community in Manchester. *Journal of Development Studies*, *47*(6), 894-912.

Asaad, Z., & Marane, B. (2020). Corruption, Terrorism and the Stock Market: The Evidence from Iraq. *The Journal of Asian Finance, Economics and Business (JAFEB)*, *7*(10), 629-639.

Aydas, O. T., Metin-Ozcan, K., & Neyapti, B. (2005). Determinants of workers' remittances: the case of Turkey. *Emerging Markets Finance and Trade*, *41*(3), 53-69.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*(1), 135-171.

Bai, J., & Liao, Y. (2016). Efficient estimation of approximate factor models via penalized maximum likelihood. *Journal of Econometrics*, *191*(1), 1-18.

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. Econometrica, 70(1), 191–221.

Bai, J., & Ng, S. (2006a). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. Econometrica, 74(4), 1133–1150.

Bai, J., & Ng, S. (2006b). Evaluating latent and observed factors in macroeconomics and finance. Journal of Econometrics, 131(1–2), 507–537.

Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. Journal of Econometrics, 146(2), 304–317.

Bai, J., & Ng, S. (2009). Boosting diffusion indices. *Journal of Applied Econometrics*, *24*(4), 607-629.

Banerjee, A., & Marcellino, M. (2008). Factor-augmented error correction models. CEPR Discussion Papers 6707, C.E.P.R. Discussion Papers.

Barrodale, I., & Roberts, F. D. (1973). An improved algorithm for discrete l_1 linear approximation. *SIAM Journal on Numerical Analysis*, *10*(5), 839-848.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*(2), 521-547.

Bernanke, B. S., Boivin, J., & Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly journal of economics*, *120*(1), 387-422.

Bernanke, B.S., Boivin, J., 2003. Monetary policy in a data-rich environment. *Journal of Monetary Economics 50*, 525–546.

Black, B. S., & Gilson, R. J. (1998). Venture capital and the structure of capital markets: banks versus stock markets. *Journal of financial economics*, *47*(3), 243-277.

Boivin, J., Ng, S., 2005. Undertanding and comparing factor based forecasts. *International Journal of Central Banking 1(3),* 117-152.

Boivin, J., Ng, S., 2006. Are more data always better for factor analysis? *Journal of Econometrics 132(1),* 169–194.

Breaux, H. J. 1967. On Stepwise Multiple Linear Regression. Technical Report. Aberdeen: Army Ballistic Research Lab Aberdeen Proving Ground MD.

Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, *5*(1), 232.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, *37*(4), 373-384.

Campos, J., Hendry, D. F., and Krolzig, H.-M. (2003). Consistent model selection by an automatic Gets approach. *Oxford Bulletin of Economics and Statistics*, *65,* 803–819.

Cantoni, E., Flemming, J. M., & Ronchetti, E. (2011). Variable selection in additive models by non-negative garrote. *Statistical modelling*, *11*(3), 237-252.

Castle, J. L., Clements, M. P., & Hendry, D. F. (2013). Forecasting by factors, by variables, by both or neither?. *Journal of Econometrics*, *177*(2), 305-319.

Castle, J. L., Doornik, J. A., & Hendry, D. F. (2021). Modelling non-stationary 'big data'. *International Journal of Forecasting*, *37*(4), 1556-1575.

Chiaraah, A. N. T. H. O. N. Y., & Nkegbe, P. K. (2014). GDP growth, money growth, exchange rate and inflation in Ghana. *Journal of Contemporary Issues in Business Research*, *3*(2), 75-87.

Cunha, R., Kock, A. B., & Pereira, P. L. V. (2019). Forecasting large covariance matrices: comparing autometrics and LASSOVAR.

Darne, O., & Charles, A. (2020). Nowcasting GDP growth using data reduction methods: Evidence for the French economy. *Economics Bulletin*, *40*(3), 2431-2439.

Desboulets, L. D. D. (2018). A review on variable selection in regression analysis. *Econometrics*, *6*(4), 45.

Diebold, F. X., & Shin, M. (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives. *International Journal of Forecasting*, *35*(4), 1679-1691.

Ding, A. A., & Hwang, J. T. G. (1999). Prediction intervals, factor analysis models, and high-dimensional empirical linear prediction. *Journal of the American Statistical Association*, 94(446), 446–455.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., & Picard, D. (1995). Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(2), 301-337.

Doornik, J. A., & Hendry, D. F. (2015). Statistical model selection with big data. *Cogent Economics* and Finance, http://dx.doi.org/10.1080/ 23322039.2015.1045216.

Doornik, J.A., 2009b. Econometric model selection with more variables than observations. Working paper. Economics Department, University of Oxford.

Dufour, J.-M., & Stevanovic, D. (2010). Factor-augmented VARMA models: Identification, estimation, forecasting and impulse responses. Working paper, McGill University.

Eisenstein, E. M., & Lodish, L. M. (2002). Marketing decision support and intelligent systems: precisely worthwhile or vaguely worthless. *Handbook of marketing. London, UK: SAGE*, 436-454.

Eita, J. H. (2012). Modelling macroeconomic determinants of stock market prices: Evidence from Namibia. *Journal of Applied Business Research (JABR), 28*(5), 871-884.

El-Sakka, M. I., & McNabb, R. (1999). The macroeconomic determinants of emigrant remittances. *World development, 27*(8), 1493-1502.

Epprecht, C., Guegan, D., Veiga, Á., & Correa da Rosa, J. (2019). Variable selection and forecasting via automated methods for linear models: LASSO/adaLASSO and Autometrics. *Communications in Statistics-Simulation and Computation*, 1-20.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association, 96,* 1348-1360.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70*(5), 849-911.

Fan, J., Ke, Y., & Liao, Y. (2016a). Robust factor models with explanatory proxies. *arXiv preprint arXiv:1603.07041*.

Fan, J., Liao, Y., & Wang, W. (2016b). Projected principal component analysis in factor models. *Annals of statistics, 44*(1), 219.

Fan, J., Ma, Y., & Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association, 109*(507), 1270-1284.

Fan, J., Xue, L., & Yao, J. (2017). Sufficient forecasting using factor models. *Journal of econometrics, 201*(2), 292-306.

Faust, J., & Whiteman, C. H. (1997, December). General-to-specific procedures for fitting a data-admissible, theory-inspired, congruent, parsimonious, encompassing, weakly-exogenous, identified, structural model to the DGP: A translation and critique. In *Carnegie-Rochester Conference Series on Public Policy* (Vol. 47, pp. 121-161). North-Holland.

Filzmoser, P., & Nordhausen, K. (2021). Robust linear regression for high-dimensional data: An overview. *Wiley Interdisciplinary Reviews: Computational Statistics*, *13*(4), e1524.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics*, *82*(4), 540-554.

Gavin, W. T., & Kliesen, K. L. (2006). Forecasting inflation and output: comparing data-rich models with simple rules. *FRB of St. Louis Working Paper No*.

Gijbels, I., & Vrinssen, I. (2015). Robust nonnegative garrote variable selection in linear regression. *Computational Statistics & Data Analysis*, *85*, 1-22.

Guerard, J., Thomakos, D., & Kyriazi, F. (2020). Automatic time series modeling and forecasting: A replication case study of forecasting real GDP, the unemployment rate and the impact of leading economic indicators. *Cogent Economics & Finance*, *8*(1), 1759483.

Gujarati, D. N., Porter, D. C., & Gunasekar, S. (2012). *Basic econometrics*. Tata McGraw-Hill Education.

Gupta, P. (2005). Macroeconomic determinants of remittances: Evidence from India. In IMF working papers 05/004. International Monetary Fund.

Hansen, C., & Liao, Y. (2019). The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications. *Econometric Theory*, *35*(3), 465-509.

Hendry, D. F., & Krolzig, H. M. (2005). The properties of automatic Gets modelling. *The Economic Journal*, *115*(502), C32-C61.

Hoerl, Arthur E, and Kennard, Robert W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Hoover, K. D., & Perez, S. J. (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *The econometrics journal*, *2*(2), 167-191.

Imimole, B., & Enoma, A. (2011). Exchange rate depreciation and inflation in Nigeria (1986–2008). *Business and Economics Journal*, *28*(1), 1-11.

Jaffri, A. A., Mirza, F. M., & Bashir, S. (2014). Is Passthrough of global food inflation to food inflation in Pakistan symmetric? *Pakistan Economic and Social Review*, 35-43.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Jiranyakul, K. (2019). Oil price shocks and domestic inflation in Thailand. *Available at SSRN 2578836*.

Kayamo, S. E. (2021). Asymmetric impact of real exchange rate on inflation in Ethiopia: a non-linear ARDL approach. *Cogent Economics & Finance*, *9*(1), 1986931.

Khan, F., Urooj, A., & Muhammadullah, S. AN ARIMA-ANN HYBRID MODEL FOR MONTHLY GOLD PRICE FORECASTING: EMPIRICAL EVIDENCE FROM PAKISTAN. *Pakistan Economic Review*, 2021. 4:1 (Winter 2021), PP. 61-75.

Khan, I., (2022). MODEL SELECTION PROCEDURES COMPARISON AND EVALUATION THROUGH MONTE CARLO EXPERIMENT. (PhD, Thesis)

Khan, M. S., & Schimmelpfennig, A. (2006). Inflation in Pakistan. *The Pakistan development review*, 185-202.

Kim, H. H., & Swanson, N. R. (2014a). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. Journal of Econometrics, 178(2), 352–367.

Kim, H. H., & Swanson, N. R. (2018a). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, *34*(2), 339-354.

Kim, H. H., & Swanson, N. R. (2018b). Methods for backcasting, nowcasting and forecasting using factor-MIDAS: With an application to Korean GDP. *Journal of Forecasting*, *37*(3), 281-302.

Kim, H., & Ko, K. (2020). Improving forecast accuracy of financial vulnerability: PLS factor model approach. *Economic Modelling*, *88*, 341-355.

Kim, H., & Shi, W. (2021). Forecasting financial vulnerability in the USA: A factor model approach. *Journal of Forecasting*, *40*(3), 439-457.

Kim, H., Shi, W., & Kim, H. H. (2020). Forecasting financial stress indices in Korea: a factor model approach. *Empirical Economics*, *59*(6), 2859-2898.

Kim, Y., & Kwon, S. (2012). Global optimality of nonconvex penalized estimators. *Biometrika*, *99*(2), 315-325.

Kock, A. B., & Teräsvirta, T. (2014). Forecasting performances of three automated modelling techniques during the economic crisis 2007–2009. *International Journal of Forecasting*, *30*(3), 616-631.

Kristensen, J. T. (2017). Diffusion indexes with sparse loadings. *Journal of Business & Economic Statistics*, *35*(3), 434-451.

Krolzig, H. M., & Hendry, D. F. (2001). Computer automation of general-to-specific model selection procedures. *Journal of Economic Dynamics and Control*, *25*(6-7), 831-866.

Laniran, T. J., & Adeniyi, D. A. (2015). An evaluation of the determinants of remittances: Evidence from Nigeria. *African Human Mobility Review, 1*(2), 179-203.

Leamer, E. E. (1983). Model choice and specification analysis. *Handbook of econometrics, 1*, 285-330.

Leamer, E. E., & Leamer, E. E. (1978). Specification searches: Ad hoc inference with nonexperimental data (Vol. 53). John Wiley & Sons Incorporated.

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.

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.

Lianos, T. P. (1997). Factors determining migrant remittances: The case of Greece. *International Migration Review*, *31*(1), 72-87.

Liu, H. Y., & Chen, X. L. (2017). The imported price, inflation and exchange rate pass-through in China. *Cogent Economics & Finance*, *5*(1), 1279814.

Lovell, M. C. (1983). Data mining. *Review of Economics and Statistics, 65,* 1–12.

Luciani, M. (2014). Forecasting with approximate dynamic factor models: the role of non-pervasive shocks. *International Journal of Forecasting*, *30*(1), 20-29.

Lueth, E. & Ruiz-Arranz, M. (2006). A gravity model of workers' remittances. IMF Working Paper 06/290. Washington, DC.

Maehashi, K., & Shintani, M. (2020). Macroeconomic forecasting using factor models and machine learning: an application to Japan. *Journal of the Japanese and International Economies*, *58*, 101104.

Maku O.E., and Atanda, A.A. "Determinants of Stock Market Performance in Nigeria: Long-Run Analysis." Journal of Management and Organizational Behaviour, 2010, 1, 1–16.

Mansor, R., Zulkifli, M., Yusof, M. M., Ismail, M. I., Ismail, S., & Yin, Y. C. (2014, December). Performance of fuzzy approach in Malaysia short-term electricity load forecasting. In *AIP Conference Proceedings* (Vol. 1635, No. 1, pp. 817-824). American Institute of Physics.

Mbongo, J. E., Mutasa, F., & Msigwa, R. E. (2014). The effects of money supply on inflation in Tanzania. *Economics*, *3*(2), 19-26.

Mohanty, D., & John, J. (2015). Determinants of inflation in India. *Journal of Asian Economics*, *36*, 86-96.

Muhammadullah, S., Urooj, A., Khan, F., Alshahrani, M. N., Alqawba, M., & Al-Marzouki, S. (2022). Comparison of Weighted Lag Adaptive LASSO with Autometrics for Covariate Selection and Forecasting Using Time-Series Data. *Complexity*, *2022*.

Mustafa, K., & Ali, S. R. (2018). The macroeconomic determinants of remittances in Pakistan. *International Journal of Business Management and Finance Research*, *1*(1), 1-8.

Narayan, P. K., Devpura, N., & Wang, H. (2020). Japanese currency and stock market—What happened during the COVID-19 pandemic? *Economic Analysis and Policy*, *68*, 191-198.

Nisa, M. U., & Nishat, M. (2011). The determinants of stock prices in Pakistan. *Asian Economic and Financial Review*, *1*(4), 276-291.

Odinakachi Njoku, C., & Emmanuel Nwaimo, C. (2019). The Impact of exchange rates on inflation in Nigeria (1981-2015). *Management Studies and Economic Systems*, *4*(3), 171-195.

Okoye, L. U., Olokoyo, F. O., Ezeji, F. N., OKOH, J. I., & Evbuomwan, G. O. (2019). Determinants of behavior of inflation rate in Nigeria. *Investment Management and Financial Innovations*, *16*(21), 25-36.

Papapetrou, E. (2001). Oil price shocks, stock market, economic activity and employment in Greece. *Energy economics*, *23*(5), 511-532.

Pedersen, H., & Swanson, N. R. (2019). A survey of dynamic Nelson-Siegel models, diffusion indexes, and big data methods for predicting interest rates. *Quantatitive Finance and Economics*, *3*, 22-45.

Peña, D., & Poncela, P. (2004). Forecasting with nonstationary dynamic factor models. *Journal of Econometrics*, *119*(2), 291-321.

Pretis, F., Reade, J. J., & Sucarrat, G. (2018). Automated general-to-specific (GETS) regression modeling and indicator saturation for outliers and structural breaks. *Journal of Statistical Software*, *86*, 1-44.

Pretis, F., Reade, J. J., & Sucarrat, G. (2018). Automated general-to-specific (GETS) regression

Qayyum, A. (2006). Money, inflation, and growth in Pakistan. *The Pakistan development review*, 203-212.

Qi, M., & Zhang, G. P. (2001). An investigation of model selection criteria for neural network time series forecasting. *European Journal of Operational Research*, *132*(3), 666-680.

Rapach, D. E., Strauss, J. K., & Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, *23*(2), 821-862.

Richardson, A., van Florenstein Mulder, T., & Vehbi, T. (2021). Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, *37*(2), 941-948.

Reichlin, L., Giannone, D., & Sala, L. (2005). *Monetary policy in real time* (No. 2013/10177). ULB--Universite Libre de Bruxelles.

Ricketts, J. R. (2011). Impact of macroeconomic shocks on remittance inflows to Jamaica: A VECM approach. Bank of Jamaica Working Paper.

López-Robles, J. R., Rodríguez-Salvador, M., Gamboa-Rosales, N. K., Ramirez-Rosales, S., & Cobo, M. J. (2019). The last five years of Big Data Research in Economics, Econometrics and Finance: Identification and conceptual analysis. *Procedia computer science*, *162*, 729-736.

Rocha, J. V., & Pereira, P. L. V. (2019). Automated model selection with applications to Brazilian Industrial Production index.

Saeed, M. M. (2017). Impact of political stability, government effectiveness and corruption on stock markets of South Asia. *Journal of the Punjab University Historical Society*, 30(1), 325–351.

Shah, M. A. A., Aleem, M., & Arshed, N. (2014). Statistical analysis of the factors affecting inflation in Pakistan. *Middle-East Journal of Scientific Research*, *21*(1), 181-189.

Shahbaz, M. (2013). Linkages between inflation, economic growth and terrorism in Pakistan. *Economic modelling*, *32*, 496-506.

Shahbaz, M., Tiwari, A. K., & Tahir, M. I. (2015). Analyzing time–frequency relationship between oil price and exchange rate in Pakistan through wavelets. *Journal of Applied Statistics*, *42*(4), 690-704.

Schumacher, C., & Breitung, J. (2008). Real-time forecasting of German GDP based on a large factor model with monthly and quarterly data. *International Journal of Forecasting*, *24*(3), 386-398.

Shrestha, P. K., & Subedi, B. R. (2014). Determinants of stock market performance in Nepal. *NRB Economic Review*, *26*(2), 25-40.

Smeekes, S., & Wijler, E. (2018). Macroeconomic forecasting using penalized regression methods. *International journal of forecasting*, *34*(3), 408-430.

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

Stock, J. H., & Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, *20*(2), 147-162.

Stock, J. H., & Watson, M. W. (2005). Implications of dynamic factor models for VAR analysis. NBER Working Papers 11467, National Bureau of Economic Research, Inc.

Stock, J. H., & Watson, M. W. (2006). Forecasting with many predictors. In G. Elliott, C. Granger, & A. Timmermann (Eds.), Handbook of economic forecasting, Vol. 1 (pp. 515–554). Elsevier, (chapter 10).

Stock, J.H., Watson, M.W., 1999. Forecasting inflation. *Journal of Monetary Economics 44*, 293–335.

Stock, J.H., Watson, M.W., 2012. Generalized shrinkage methods for forecasting using many predictors. *Journal of Business and Economic Statistics 30*, 481–493.

Swanson, N. R., & Xiong, W. (2018a). Big data analytics in economics: What have we learned so far, and where should we go from here? *Canadian Journal of Economics, 51*(3), 695–746.

Swanson, N. R., Xiong, W., & Yang, X. (2020). Predicting interest rates using shrinkage methods, real-time diffusion indexes, and model combinations. *Journal of Applied Econometrics, 35*(5), 587-613.

Swanson, N.R. and Xiong, W. (2018b), Predicting Interest Rates Using Shrinkage Methods, Real-Time Diffusion Indexes, and Model Combinations, Working Paper, Rutgers University.

Syed, A. A. S., & Lee, K. H. (2021). Macroeconomic forecasting for Pakistan in a data-rich environment. *Applied Economics*, *53*(9), 1077-1091.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society.* Series B (Methodological), 267-288.

Tsaurai, K. (2018). What are the determinants of stock market development in emerging markets? *Academy of Accounting and Financial Studies Journal*, *22*(2), 1-11.

Tu, Y., & Lee, T. H. (2019). Forecasting using supervised factor models. *Journal of Management Science and Engineering*, *4*(1), 12-27.

Ullah, I., Rahman, M. U., & Jebran, K. (2015). Terrorism and worker's remittances in Pakistan. *Journal of Business Studies Quarterly*, *6*(3), 178.

Vargas-Silva, C., & Huang, P. (2006). Macroeconomic determinants of workers' remittances: Hostversus home country's economic conditions. *Journal of International Trade & Economic Development*, *15*(1), 81-99.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3-28.

Wahid, A., Khan, D. M., & Hussain, I. (2017). Robust Adaptive Lasso method for parameter's estimation and variable selection in high-dimensional sparse models. *PLoS one*, *12*(8), e0183518.

Westerlund, J., Urbain, J. P., & Bonilla, J. (2014). Application of air quality combination forecasting to Bogota. *Atmospheric Environment*, *89*, 22-28.

Wold, H., 1982. Soft Modelling: The Basic Design and Some Extensions, Vol. 1 of Systems under Indirect Observation, Part II. North-Holland, Amsterdam.

Xiao, N., and Xu, Q. S. (2015). Multi-step Elastic-net: reducing false positives in high dimensional variable selection. *Journal of Statistical Computation and Simulation, 85*(18), 3755-3765.

Zeng, L., and Xie, J. (2014). Group variable selection via SCAD-L 2. *Statistics*, *48*(1), 49-66.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, *38*(2), 894-942.

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.

Zou, H., and Zhang, H. H. (2009). On the Adaptive Elastic-net with a diverging number of parameters. *Annals of statistics, 37*(4), 1733.

Appendix A

Table A1 describes the variables, symbols, definition of each variable and source of data.

Table A1. Variables description

Sr. No	Variables	Symbol	Definition/Construction	Source of Data
1	**Workers' Remittances**	WR	The transfer of foreign money by migrated workers to Pakistan.	SBP
2	**Interest Rate**	INT	Call money rate	SBP
3	**Gold prices**	GOLD	Gold prices is defining the price of gold in which the gold is traded on gold market.	SBP
4	**Development expenditure**	DEX	It is the type of expenditure which helps economic and social development on the country. For example, the expenditure on education, health etc.	SBP
5	**Major agriculture crops**	AGC	Major agriculture crops are wheat, rice, cotton, sugarcane, maize etc.	SBP
6	**Inflation**	INF	Inflation is the increase in price of goods and services over time in general level. Inflation rate is measured by $CPI_t - CPI_{t-1} / CPI_{t-1} * 100$	SBP
7	**Foreign direct investment**	FDI	FDI is the type of investment in which the people or organization of one country invested in company of property of other countries.	SBP
8	**Trade openness**	TO	Trade openness is defined as the ratio of trade to GDP.	SBP

9	**Exchange rate/Nominal exchange rate**	EXR	Value of the rupees per unit of US dollar	IFS
10	**Stock market performance**	SP	Share prices	IFS
11	**Investment return of Pak**	IRPak	$0.8INT_{Pk} + 0.2dLn\ (SP_{Pk})$ Where INT_{Pk} is interest rate and SP_{Pk} is share prices of Pakistan.	IFS
12	**Investment return of US**	IRUS	$0.8INT_{US} + 0.2dLn\ (SP_{US})$ Where INT_{US} is interest rate and SP_{US} is share prices of US.	IFS
13	**Real Domestic Product**	GDP	It is defined as, the total value of final goods and services which are produced inside the boundary of the country in a given period.	WDI
14	**Unemployment**	UEMP	Unemployment is defined as, the people who want to work but do not have a job.	WDI
15	**Foreign debts**	DEBT	Foreign debt is a money that one country borrowed from outside country or organization. It is also known as external debt.	WDI
16	**Real effective exchange rate**	REER	It is defined as, the nominal effective exchange rate which is divided by a price deflator.	WDI
17	**Secondary school enrolment**	SSEN	Secondary school enrolment is defined as the number of students which are enrolled in secondary school.	WDI

18	**Financial Liberalization**	FINL	The data on **f**inancial liberalization is taken from Shabbir (2013). He used the following formula for the construction of financial liberalization.	Shabbir (2013)
19	**Job skill index**		The Job skill index is constructed with the help of weighted index of the different skill categories.	Bureau of Emigration and Overseas Employment
20	**Wage rate**	WAGE	The amount of wage that is paid to the worker per unit of time.	Bhatti(2018)
21	**Democracy**	DMOC	Democracy is the type of government in which people elect their representatives.	ICRG
22	**Internal Conflict**	ICNF	Internal conflict is defined as, the political violence inside the country and its actual influence on the governance.	ICRG
23	**External Conflict**	XCNF	External conflict is defined as, the problem such as; diplomatic pressures, trade restrictions etc. to the mandatory government from the foreign action to violent external pressure.	ICRG
24	**Law and order situation**	LAOR	Law and order situation is defined as the condition when people follow the rule and regulation. There is no violence or threats, and the police control all the crime etc.	ICRG
25	**Corruption**	CRRP	The illegal actions by powerful people such as bureaucrats, government, police etc.	ICRG

26	**Terrorism index (no' of attacks)**	TIND	It is the use of violence and threats for the purpose of achieving political and ideological objectives.	ICRG
27	**Government stability**	GS	Whenever the representative of the govt. change without any threats of violence it is known as political stability.	ICRG
28	**Black Market Premium**	BMP	Black market premium is defined as, the percentage difference between the black market exchange rate and official exchange rate.	ICRG
29	**Gross fixed capital formation**	GFCF	The physical capital is measured by gross fixed capital formation (GFCF). Theoretically, the relation between economic growth and capital formation is described by "Q" theory.	WDI
30	**Money Supply**	MS	The money supply is the total amount of money in circulation in a country or group of countries in a monetary union.	WDI
31	**Oil prices**	OP	It is taken as crude oil (per barrel).	IFS

Table A2. Variables Description

Sr. no	T	Name of the Variables
	4	**Real Sector (Output)**
1	4	Production of Sugar (SA)
2	4	Production of Vegetable (SA)
3	4	Production of Cigarettes (SA)
4	4	Production of Cotton yarn (SA)
5	4	Production of Cotton Cloth (SA)
6	4	Production of Paper (SA)
7	4	Production of Paper Board (SA)
8	4	Production of Soda Ash (SA)
9	4	Production of Caustic Soda (SA)
10	4	Production of Sulfuric Acid (SA)
11	4	Production of Chlorine Gas (SA)
12	4	Production of Urea (SA)
13	4	Production of Super Phosphate (SA)
14	4	Production of Ammonium Nitrate (SA)
15	4	Production of Nitro Phosphate (SA)
16	4	Production of Cycle Tyres & Tubes (SA)
17	4	Production of Motor Tyres & Tubes (SA)
18	4	Production of Cement (SA)

19	4	Production of Tractors (SA)
20	4	Production of Bicycle (SA)
21	4	Production of Silica Sand (SA)
22	4	Production of Gypsum (SA)
23	4	Production of Limestone (SA)
24	4	Production of Rock Salt (SA)
25	4	Production of Coal (SA)
26	4	Production of Chromate (SA)
27	4	Production of Crude Oil (SA)
28	4	Production of Natural Gas (SA)
29	4	Production of Electricity (SA)
		Monetary Sector (Money, Reserves and Banking System) **Money**
30	4	Currency in circulation
31	4	Bank Deposit with State Bank of Pakistan
32	4	Other Deposit with State Bank of Pakistan
33	4	Currency in Tills of Scheduled Banks
34	4	Demand Deposits
35	4	Time Deposits
36	4	Resident Foreign Currency Deposits
37	4	Government Sector Borrowing (net)

38	4	Budgetary Support
39	4	Commodity Operations
40	4	Credit to Private Sector
41	4	Credit to Public Sector Enterprises
42	4	Net Foreign (Domestic) Assets of State Bank of Pakistan
43	4	Net Foreign Assets of the Scheduled Banks in Pakistan
		Prices
44	4	Consumer Price Index
45	4	Consumer Price Index (Food)
46	4	Wholesale Price Index
47	4	Sensitive Price Index
		Exchange Rates
48	4	Nominal Effective Exchange Rate
49	4	Real Effective Exchange Rate
50	4	Saudi Arabian Riyal (Monthly Average)
51	4	UAE Dirham (Monthly Average)
52	4	US Dollar (Monthly Average)
53	4	Canadian Dollar (Monthly Average)
54	4	UK Pound Sterling (Monthly Average)
55	4	Euro (Monthly Average)
56	4	Japanese Yen (Monthly Average)

		Interest Rates
57	2	Lending Weighted Average Rates
58	2	Deposits Weighted Average Rates
59	2	Call Money Rate
60	2	Overnight Weighted Average Repo Rate (all data)
61	2	Karachi Interbank Offered Rate 1 Week
62	2	Karachi Interbank Offered Rate 2 Weeks
63	2	Karachi Interbank Offered Rate 1 Month
64	2	Karachi Interbank Offered Rate 3 Months
65	2	Karachi Interbank Offered Rate 6 Months
66	2	Karachi Interbank Offered Rate 9 Months
67	2	Karachi Interbank Offered Rate 12 Months
		External Sector
68	4	Exports
69	4	Imports
70	4	Workers Remittances
71	4	Gold Reserves
72	4	Foreign Exchange Reserves with State Bank of Pakistan
73	4	Foreign Exchange Reserves with Scheduled Banks in Pakistan
74	4	Old Foreign Currency Accounts
75	4	New Foreign Currency Accounts (FE-25)

		Fiscal Sector
76	4	Federal Government Direct Tax Collection
77	4	Federal Government Indirect Tax (Sales Tax)
78	4	Federal Government Indirect Tax (Excise Tax)
79	4	Federal Government Indirect Tax (Customs)

Appendix B

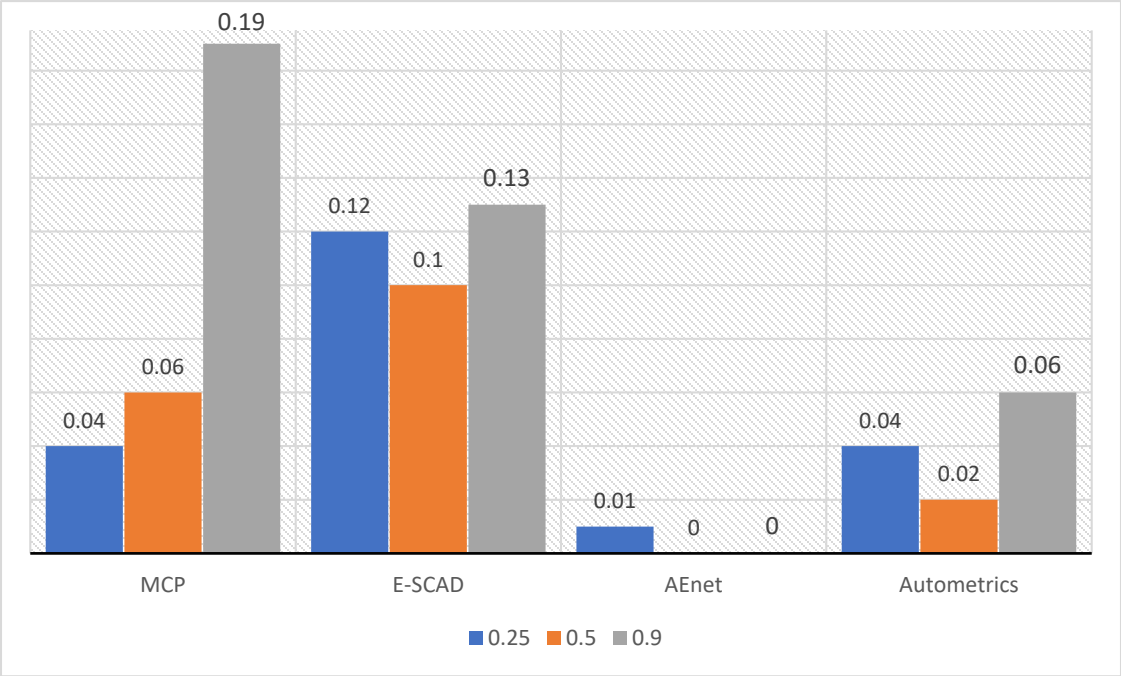
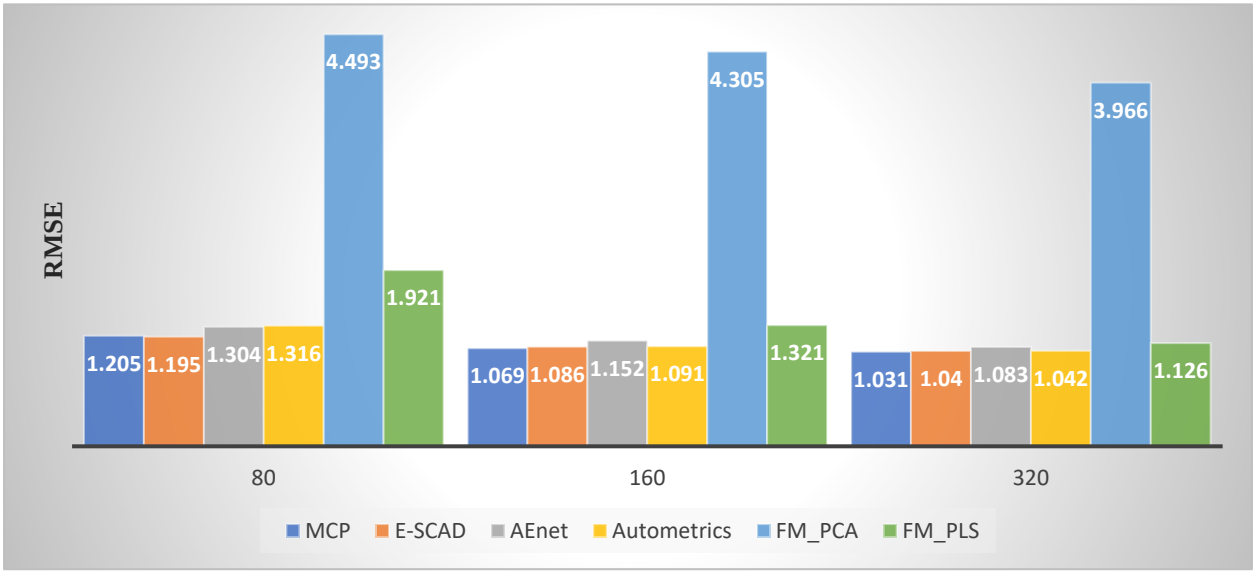
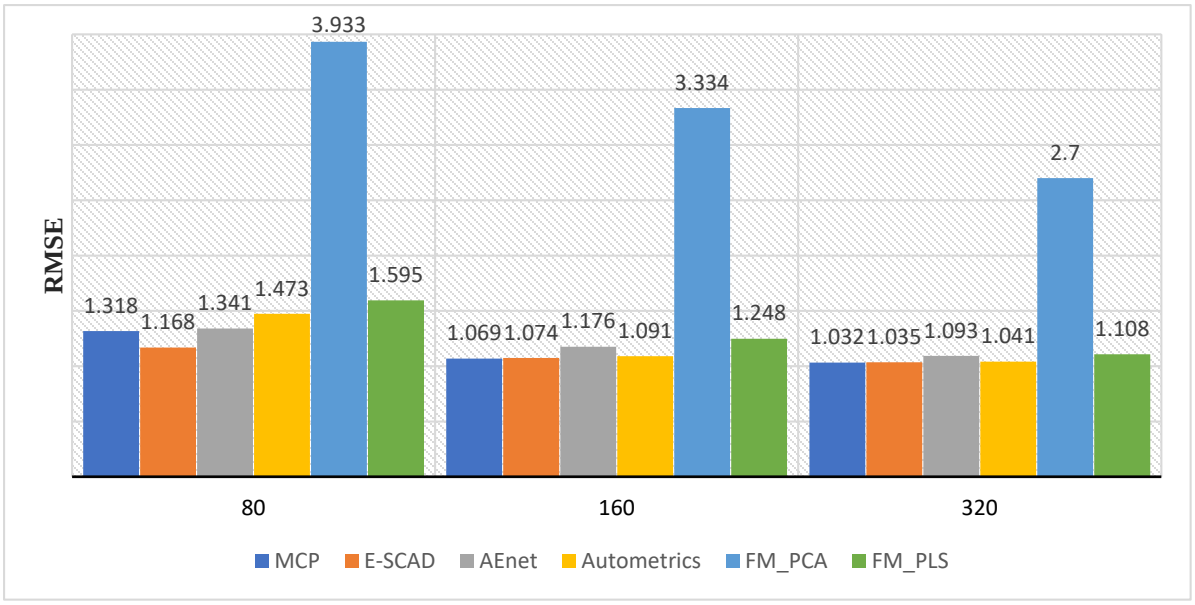


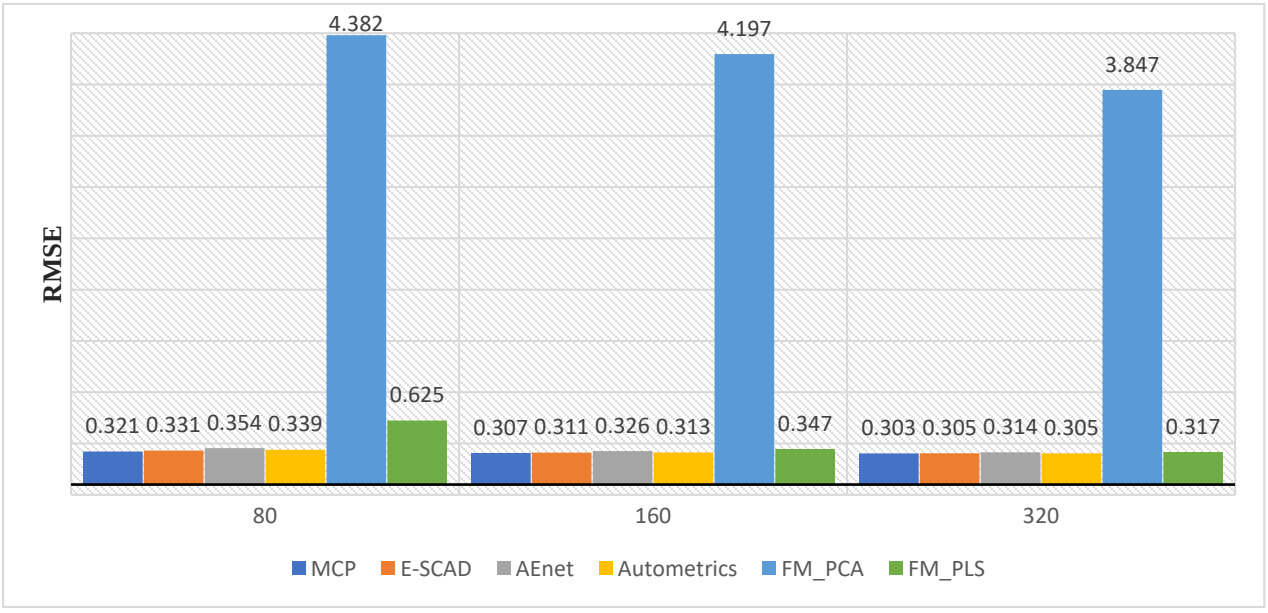
Figure 1: Computation of Gauge across three cases of multicollinearity, when n = 80 and P = 50.



(a)

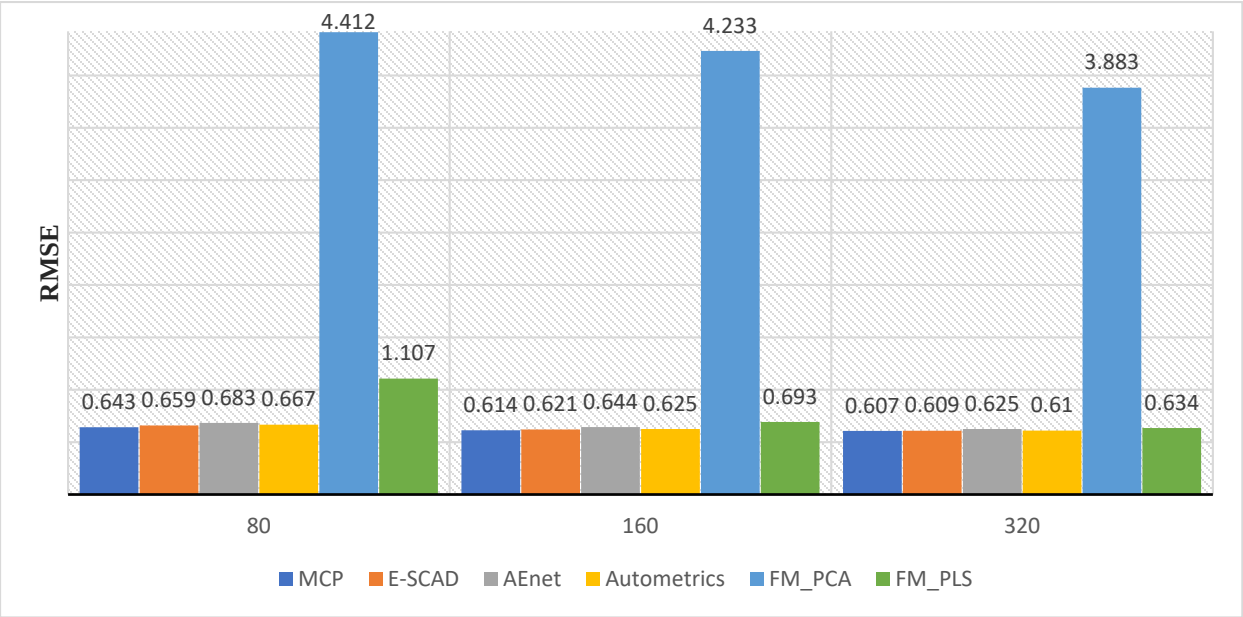

RMSE
3.933
3.334
2.7
1.318
1.168
1.341
1.473
1.595
1.069
1.074
1.176
1.091
1.248
1.032
1.035
1.093
1.041
1.108
80
160
320
MCP
E-SCAD
AEnet
Autometrics
FM_PCA
FM_PLS

(b)

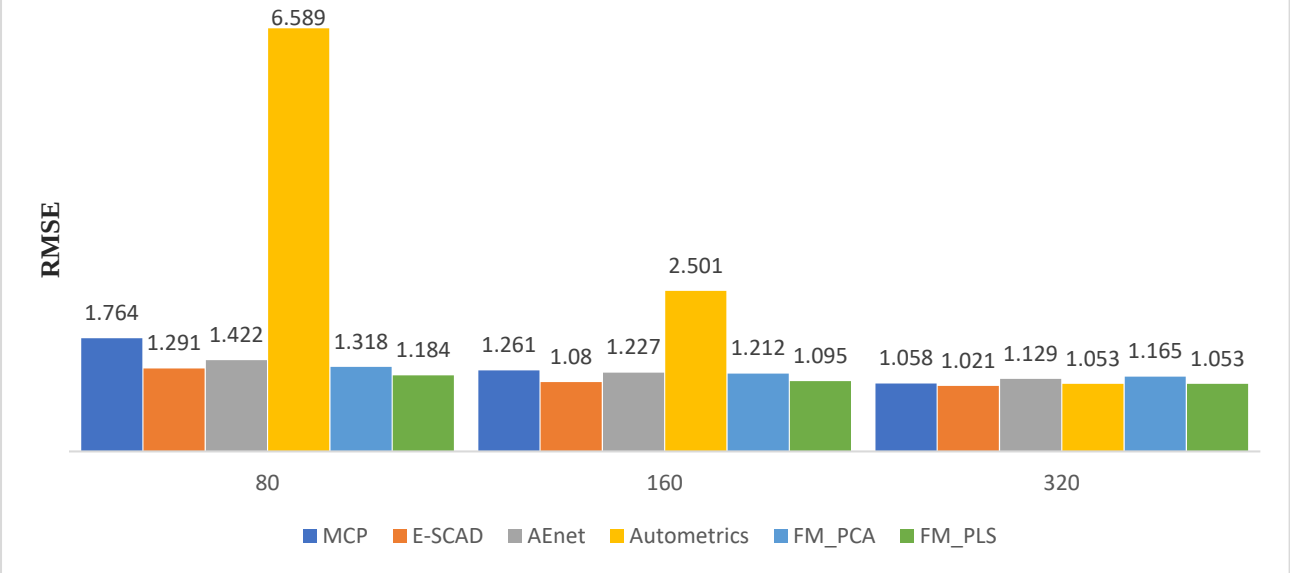

RMSE
4.382
4.197
3.847
0.321
0.331
0.354
0.339
0.625
0.307
0.311
0.326
0.313
0.347
0.303
0.305
0.314
0.305
0.317
80
160
320
MCP
E-SCAD
AEnet
Autometrics
FM_PCA
FM_PLS

(c)

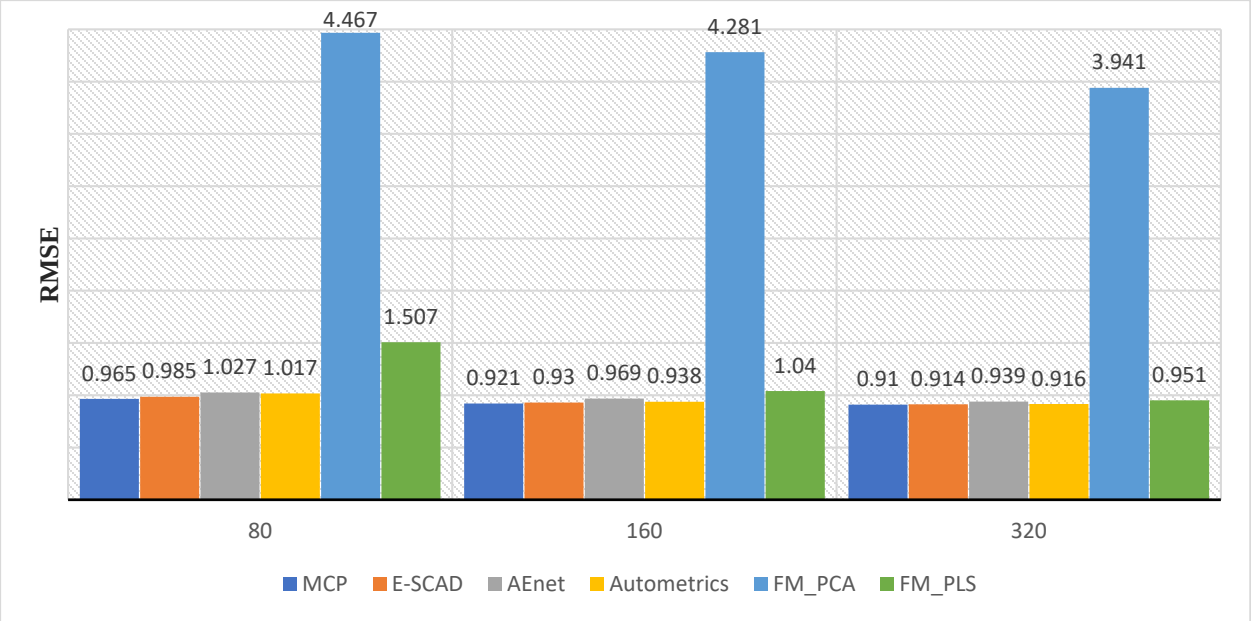
Figure 2: Out of sample root mean squares error across sample size, where forecasts are obtained from various models when ρ = 0.25(a), ρ = 0.5(b), ρ = 0.90(c), and P = 70.



(a)

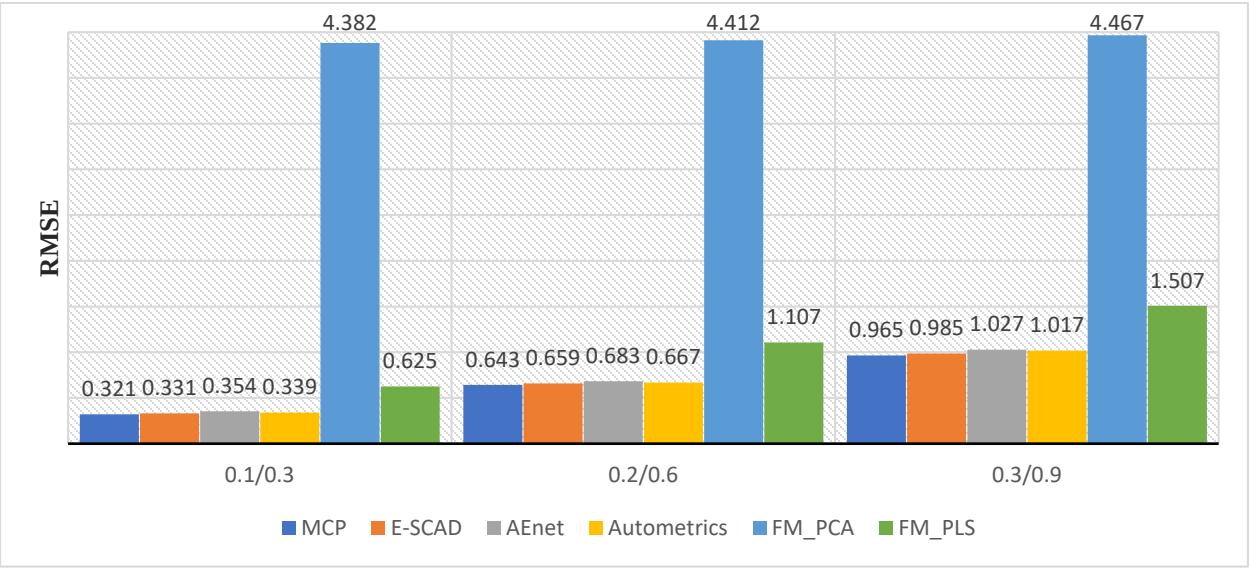


(b)

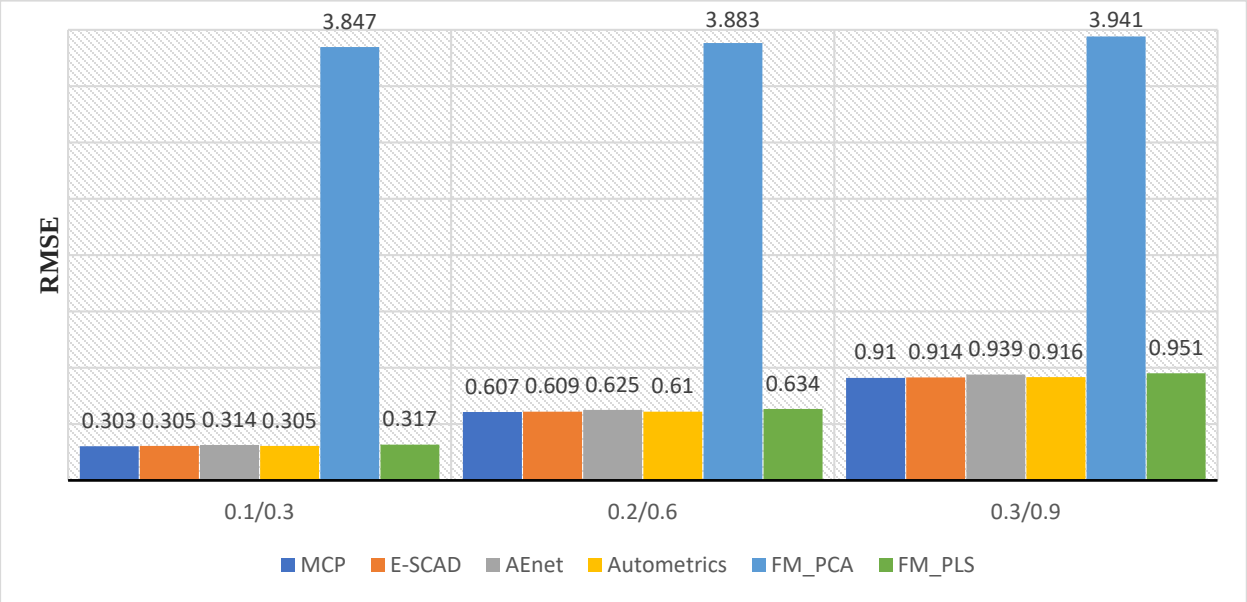


(c)

Figure 3: Out of sample root mean squares error across sample size, where forecasts are achieved from various models when π_1 = 0.1/0.3(a), π_2 = 0.2/0.6(b), π_3 = 0.3/0.9(c) and P = 70.



(a)



(b)

Figure 4: Out of sample root mean squares error across the levels of Heteroscedasticity, where forecasts are achieved from various models when n = 80(a), n = 320(b), and P = 70.

Appendix C

Package Details

In this study, various methods are used. We have used the R version 4.2 and 4.3 in our study.

Autometrics: We use GETS (general to specific) package in R, in order to estimate the model, which retains the relevant variable (also called parsimonious model). After estimating the model, we did not use the forecast package for forecasting purpose, rather we design the code and achieve the forecast. The details are given below:

- gets: General-to-Specific (GETS) Modelling and Indicator Saturation Methods

Author: Genaro Sucarrat [aut, cre], Felix Pretis [aut], James Reade [aut], Jonas Kurle [ctb], Moritz Schwarz [ctb]

Maintainer: Genaro Sucarrat <genaro.sucarrat at bi.no>

Pretis, Reade and Sucarrat (2018) <doi:10.18637/jss.v086.i03>.

Shrinkage Methods: Similarly, for estimating the Elastic SCAD and MCP models, we use 'ncvreg' package. For adaptive Elastic net, we glmnet package. Glmnet package is unable to estimate the adaptive elastic net package directly, thus we made some adjustment in the codes for achieving the output from adaptive elastic net. For forecasting, we forecast package.

Author: Patrick Breheny [aut, cre]

Maintainer: Patrick Breheny <patrick-breheny at uiowa.edu>

- ncvreg: Regularization Paths for SCAD and MCP Penalized Regression Models

Breheny and Huang (2011) <doi:10.1214/10-AOAS388> or visit the ncvreg homepage <https://pbreheny.github.io/ncvreg/>.

Author: Jerome Friedman [aut], Trevor Hastie [aut, cre], Rob Tibshirani [aut], Balasubramanian Narasimhan [aut], Kenneth Tay [aut], Noah Simon [aut], Junyang Qian [ctb], James Yang [aut]

Maintainer: Trevor Hastie <hastie at stanford.edu>

- glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models

URL: https://glmnet.stanford.edu, https://dx.doi.org/10.18637/jss.v033.i01, https://dx.doi.org/10.18637/jss.v039.i05

Factor Models: For factor models, we use caret and pls packages. Modified the codes accordingly.

- caret: Classification and Regression Training

Author: Max Kuhn [aut, cre], Jed Wing [ctb], Steve Weston [ctb], Andre Williams [ctb], Chris Keefer [ctb], Allan Engelhardt [ctb], Tony Cooper [ctb], Zachary Mayer [ctb], Brenton Kenkel [ctb], R Core Team [ctb], Michael Benesty [ctb], Reynald Lescarbeau [ctb], Andrew Ziem [ctb], Luca Scrucca [ctb], Yuan Tang [ctb], Can Candan [ctb], Tyler Hunt [ctb]

Maintainer: Max Kuhn <mxkuhn at gmail.com>

https://github.com/topepo/caret/issues

- pls: Partial Least Squares and Principal Component Regression

Author: Kristian Hovde Liland [aut, cre], Bjørn-Helge Mevik [aut], Ron Wehrens [aut], Paul Hiemstra [ctb]

Maintainer: Kristian Hovde Liland <kristian.liland at nmbu.no>

https://github.com/khliland/pls/issues