Erasmus University Rotterdam
Erasmus School of Economics
Seminar Machine Learning — Econometrics and Economics

Examining the Efficacy of Machine Learning Techniques in Macroeconomic Forecasting for Developing Economies

Armand Bonn, Isabella de Graaff, Hicham Kamhi, and Mia Sin

Abstract

In the realm of macroeconomic forecasting, the adaptation of machine learning techniques to developing economies offers a promising frontier for research. Machine learning techniques have been shown to provide new insights and often have a better performance than traditional methods. This study compares the performance of the Adaptive Lasso and Sparse Principal Component (SPC) methodologies under conditions of limited data availability and varying economic contexts in Brazil and Costa Rica, respectively. These countries are distinctly different in their economic profiles. Utilizing a dataset comprising thirty macroeconomic indicators to forecast Real Gross Domestic Product ($RGDP$) and Consumer Price Index (CPI), this paper examines whether penalized regression methods outperform traditional factor models in developing countries, as recent evidence from US data suggests. The findings reveal mixed results, highlighting that while Adaptive Lasso shows superior performance in Brazil for CPI, traditional factor models remain competitive in Costa Rica. This suggests that the optimal choice of forecasting methodology may depend on the specific economic characteristics and data environment of the respective country. However, an important conclusion of the findings is the fact that dimensionality reduction with scarce data from developing countries generally leads to poor forecasting of the dependent variable. Business cycles in the dependent variables are ineffectively forecasted for both model types. The research contributes to the literature by providing empirical evidence on the applicability of these advanced forecasting techniques in less-studied developing economic contexts and by offering insights into macroeconomic forecasting strategies that can be applied in data-poor contexts.

Supervisor:	Max Welz
Date final version:	5th December 2024

1 Introduction

Accurate forecasting of macroeconomic variables is a cornerstone of effective policy-making and devising investment strategies in developing economies, where economic conditions are frequently volatile and data are often scarce. This study delves into the application of machine learning techniques, specifically the Adaptive Lasso, against traditional factor models, specifically Sparse Principal Components, to enhance the forecasting accuracy of macroeconomic indicators. By comparing the performance of these methodologies in Brazil and Costa Rica—each representing distinct economic contexts— we assess the relative performance of these advanced methodologies against traditional econometric models, thereby contributing to the literature on economic forecasting in under-researched settings.

Conventionally, factor models have been employed in empirical applications to reduce the high number of variables often present in macroeconomic data. In their paper, Smeekes & Wijler (2018) assessed the relative performance of penalized regression techniques compared to traditional models in macroeconomic forecasting to contribute to unexplored literature on this topic. The goal of penalized regression techniques is to use a model that selects relevant variables for forecasting. Through the use of extensive datasets, applying regularization techniques can be justified and the authors show strong performance of these models. The accurate forecasting of macroeconomic indices is particularly relevant in the context of developing economies, where the prediction of economic trends can be hindered by challenges such as scarce and unreliable data. These complexities are further exacerbated in environments characterized by frequent changes in macroeconomic policies, typically triggered by factors such as economic volatility, political instability, and reactions to global economic dynamics (Chuku et al., 2019). In such settings, governments are compelled to continually adjust monetary, fiscal, and trade policies in efforts to foster economic stability and growth, which makes improving the precision of macroeconomic forecasting an important endeavor. The combination of the volatile economic environment and scarce availability of data for developing economies make it compelling to undertake further research. The primary inquiry now is to assess how these techniques perform when applied in the macroeconomic context of developing countries.

Macroeconomic forecasting in developing countries has received less attention in existing literature, particularly concerning the application of penalized regression techniques. It seems that this stems primarily from the lack of data availability, quality, and frequency, which makes it challenging to apply and validate the efficacy of forecasting methods. Despite these data challenges, this study sheds light on the efficacy of machine learning techniques in macroeconomic forecasting in environments characterized by greater economic volatility and less predictability. We focus on Brazil and Costa Rica in particular, as both countries have different factors that drive macroeconomic variables in their respective countries, which makes an interesting comparison for the applicability of penalized regression methods relative to factor models. Developing countries with small open economies are generally more sensitive to external shocks, as discussed by Leon-Gonzalez & Thu (2021), whereas developing countries with larger open economies are more robust to external shocks, while domestic shocks are the main source of, among others, GDP fluctuations (Hoffmaister & Roldos, 2001).

Smeekes & Wijler (2018) have discussed the comparative effectiveness of penalized regression

techniques and factor models for US macroeconomic data and shown promising results from these alternative dimensionality reduction techniques. One of the main findings of Smeekes & Wijler (2018) states that the traditional factor models still outperform lasso-type methods for real series, with Sparse Principal Components (SPC), which will be used interchangeably with SPCA, giving the best forecasts for these variables. On the other hand, the lasso-type methods, particularly Adaptive Lasso, show at least as good or even better performance for nominal series in the authors' empirical application. As research in the area of macroeconomic forecasting for developing economies is scarce, this paper fills part of this gap and gives a more extensive representation of the comparative performance of factor models and penalized regression models.

Our research will focus on forecasting two macroeconomic variables, namely the Real Gross Domestic Product and (Nominal) Consumer Price Index, and comparing the performance of the lasso-type models and two Principal Component type models. We focus on the Adaptive Lasso due to its improved variable selection accuracy, as this adapted version of the Lasso model seems to provide better variable selection performance, and strong performance in the empirical application of Smeekes & Wijler (2018). Furthermore, we focus on the Sparse Principal Component model because, not only has it shown strong performance in the findings of Smeekes & Wijler (2018), but also due to its robustness and parsimony. Due to significant differences in the number of observations and variables in the datasets, US data may not be representative of data in other, particularly developing or underdeveloped, countries. Therefore, it might not be the case that the application of this paper finds similar results to Smeekes & Wijler (2018). By providing a comprehensive study of the comparative forecasting performance of penalized regression methods relative to factor models on macroeconomic conditions in emerging economies, the existing literature is complemented with an additional perspective on these methods and an application to volatile economic conditions. This study seeks not only to advance academic knowledge but also to provide practical, empirically validated tools for policymakers striving to achieve economic stability and growth in an increasingly unpredictable global landscape.

Our findings contribute to the understanding of the applicability of machine learning techniques as forecasting methodologies in emerging economies. The research reveals that the Adaptive Lasso model while providing robust performance for Brazil, particularly in forecasting the Consumer Price Index, does not consistently outperform Principal Component (PC) type models, which will be used interchangeably with PCA, across all contexts. In Costa Rica, a smaller economic environment, factor models retain their competitiveness, suggesting that the optimal forecasting model may vary based on specific country characteristics and the nature of economic data. This discrepancy highlights the importance of contextual economic factors in the selection of forecasting techniques, suggesting that a one-size-fits-all approach may not be adequate for the diverse economic landscapes of developing countries. These insights pave the way for a more targeted approach to economic forecasting, emphasizing the critical role of model selection in achieving accurate economic forecasts in data-constrained environments.

The remainder of the paper is structured as follows. Section 2 provides a comprehensive overview of the existing literature, exploring the evolution and applications of factor models and penalized regression techniques within the field of macroeconomic forecasting. Section 3 describes the datasets utilized for Brazil and Costa Rica, including the selection of variables and

the data preprocessing measures undertaken to ensure robust analysis. Section 4 delves into the specifications of the Adaptive Lasso and Sparse Principal Components, detailing their theoretical foundations and implementation. Section 5 presents the empirical findings, discussing how the different models perform under the economic conditions of the countries studied and discussing implications from these results. Finally, Section 6 concludes on the insights gained from the research, discusses its limitations, and suggests directions for future research.

2 Literature review

The field of macroeconomic forecasting has a significant amount of related literature. In the past, time series methods were often utilized to predict macroeconomic trends as proposed by Granger & Newbold (1977). Stock & Watson (1999) show that these methods can often provide accurate macroeconomic forecasts but also acknowledge that time series forecasting comes with challenges due to the inherent complexity of economic dynamics. Croushore (2006) and Fildes & Stekler (2002) also used these time series models to forecast macroeconomic variables with real-time data. They came to the same conclusion that time series methods are useful for macroeconomic forecasting but emphasize the need for careful model specification, data handling, and forecast evaluation. Moreover, they concluded that forecasts from multiple models should be combined to improve overall accuracy and mitigate the limitations of the separate models.

There is an extensive literature on the topic of Principal Components model and its applications to macroeconomic models. Lawley & Maxwell (1973) give a general form for the Principal Components model and the classic factor model and derive the properties of the factor and factor loadings for the model. Forni et al. (2000) analyze the consistency of the estimates in the factor model when the leads of the factor are added to the static factor model giving rise to the more general dynamic factor model. Stock & Watson (2002a) show that the estimates of a principal component are still asymptotically consistent when introducing serial correlation in the factors and errors of the factor estimation in macroeconomic data. These results are strong and give proof that the traditional factor model can be used in time series data with more data and complex relationships between variables. Subsequently, Bai & Ng (2002) analyze how the relevant number of factors can be estimated consistently from any type of macroeconomic data. Using these results, the statistical method with newly composed selection criteria to determine the number of factors is robust and applicable to complex macroeconomic time-series data as used in this paper. Alessi et al. (2010) improve these selection criteria by introducing a tuning parameter, showing that the number of factors can be estimated correctly even with large idiosyncratic components of the variables. Artis et al. (2005) make use of the results of the two formerly discussed articles and apply the factor forecast model to UK Macroeconomic data. The authors use the Principal Components model and selection criteria of Bai & Ng (2002) to select the factors for the UK data. Their results combined with those of Stock & Watson (2002a) further affirm that the use of factor models in a macroeconomic context gives stronger results relative to the time series model.

However, one of the criticisms of Zou et al. (2006) and Smeekes & Wijler (2018) is that the Principal Components model is difficult to interpret as the components are a linear combination of all variables. Stock & Watson (2002b) provide a method to enhance the interpretability of

macroeconomic series with a large number of variables. Furthermore, Zou et al. (2006) give an analytical expression of a modified form for the Principal Components model to acquire the sparse expression giving us better interpretability of the PC model. The authors combine the mathematical form of a penalized regression model with the expression of the traditional Principal Component, namely the Sparse Principal Components model. In their research, the authors give mathematical proof of the strength and consistency of SPC, namely that it is computationally efficient and a sparse reduction of PCA, and that it is able to identify variables correctly. Kristensen (2017) further analyzes the SPC model to apprehend whether the estimators of this modified version of the PC are still asymptotically consistent. More specifically, the author shows that under the assumptions of Stock & Watson (2002a) the penalized Principal Component function converges to the asymptotic objective function in Stock & Watson (2002a). The results of Kristensen (2017) show that the use of the SPC improves forecasting accuracy relative to the PC, which supports using SPC for macroeconomic forecasting. This also explains the improved forecasting accuracy of the SPC model in the empirical applications of Smeekes & Wijler (2018).

Additionally, the Adaptive Lasso model, where adaptive weights are used for penalizing different coefficients, will be analyzed. This model was introduced by Zou (2006) as a robust extension to the Lasso model. This research shows that the oracle property holds, meaning that the Adaptive Lasso model performs as well as if the actual underlying model was known and that the model can be solved efficiently. Huang et al. (2008) come to the same conclusion on the validity of the oracle property and the computational efficiency when applying the Adaptive Lasso method to high-dimensional data. Furthermore, Medeiros & Mendes (2016) show that when the number of candidate variables is higher than the number of observations, the relevant variables are consistently chosen if the number of observations increases. Moreover, they show that the oracle property still holds for the Adaptive Lasso method if the errors are non-Gaussian and conditionally heteroskedastic. Therefore, they argue that it is justifiable to apply the Adaptive Lasso to empirical applications in financial and macroeconomic data. Medeiros & Mendes (2016) applied the Adaptive Lasso model to forecast the monthly US inflation and found that the Adaptive Lasso model has stronger forecast ability than traditional models. Smeekes & Wijler (2018) also uses the Adaptive Lasso to forecast monthly macroeconomic data in the US. They find that the Adaptive Lasso performs the best out of the five shrinkage methods discussed and factor models for nominal series.

3 Data

The research on the forecasting efficacy of Adaptive Lasso and SPC in macroeconomic context for developing economies diverges from the empirical application in Smeekes & Wijler (2018) by focusing on Brazil and Costa Rica, instead of the United States. The dataset for both countries consists of quarterly observations on 32 macroeconomic variables, spanning from 2001 to 2021, resulting in 84 observations, obtained from the OECD database and Federal Reserve of St. Louise' FRED database. Compared to the analysis performed by Smeekes & Wijler (2018), who use a dataset consisting of 133 variables with monthly observations over a span from 1959 to 2015, our research is quite limited in terms of number of variables and observations.

The difference in total observations inherently imposes certain limitations on the forecast techniques as they are generally applied in high-dimensional data settings. The forecast techniques could therefore end up deleting too many variables, or not be able to find a sufficient factor structure that underlies the data.

While Smeekes & Wijler (2018) have a significantly larger dataset for their empirical application, they also compare the prediction accuracy of, among others, lasso-type estimators and principal components using an analysis based on a simulation study. In this analysis, the authors run 1000 iterations simulating 100 observations in every iteration to predict the 1-step-ahead forecast. The accuracy of these forecasts is evaluated using the Mean Squared Forecast Error (MSFE). Compared to this, the dataset utilized in this paper does not differ too much in terms of the number observations. However, this research still recognizes the limitation of using a smaller dataset, which potentially leads to more noisy observations.

The dataset for our investigation consists of 30 macroeconomic variables aimed at forecasting two dependent variables: Real Gross Domestic Product ($RGDP$) which measures real economic activity, and Consumer Price Index (CPI) which is a price index. All these variables encompass a wide range of economic indications that are measured either in their national currency or in US dollars and provide a comprehensive overview of the economic landscape of each country. For a full overview of the variables used, see table 2 in the Appendix.

To account for the uncertainty of the presence of cointegration in the datasets, we correct all series for non-stationarity in the variables to forecast as done by Smeekes & Wijler (2018). This entails taking log differences for the real variable $RGDP$ and taking log second differences for the price index CPI. Furthermore, we normalize the independent variables using the Yeo-Johnson transformation, which transforms the data such that it closely approximates a normal distribution or stabilizes its variance Smeekes & Wijler (2018). Yeo & Johnson (2000) propose this method as an extension to the Box-Cox transformation, which allows for the transformation of data irrespective of its sign, making it particularly useful for variables that include non-positive values. The Yeo-Johnson transformation is defined as:

$$y(\phi) = \begin{cases} \frac{(y+1)^{\phi}-1}{\phi} & \text{if } \phi \neq 0 \\ \log(y+1) & \text{if } \phi = 0, \end{cases} \tag{1}$$

where ϕ determines the type of transformation that is applied to the data such that it becomes closely approximated to a normal distribution and stabilizes its variance. Therefore, it is crucial to find the right ϕ parameter for applying the Yeo-Johnson transformation effectively. The ϕ parameter that is chosen for the Yeo-Johnson transformation in this paper is the default option provided by the Yeo-Johnson transformation method in R.

4 Methodology

4.1 General macroeconomic model specification

There is a widespread formulated specification of a macroeconomic time series model with high dimensionality in the cross-section, namely, the specification used by Stock & Watson (2002a).

Also, the specification used by Smeekes & Wijler (2018) will be explained and applied to this research. The objective is to forecast an economic time series with information available up to time $t = 1, ..., T$ for h steps ahead. In order to achieve this, a pre-determined set of variables and a set of candidate variables are included. The pre-determined set of variables are variables that should be included based on macroeconomic theory and lags of the dependent variables. The candidate variables are variables that are potentially relevant to estimate the dependent variable. This gives the following general model:

$$y_{t+h} = \mathbf{w}_t'\boldsymbol{\beta}_w + \mathbf{x}_t'\boldsymbol{\beta}_x + \epsilon_{t+h}, \tag{2}$$

where y_{t+h} is the dependent variable to forecast, and h the number of steps ahead to forecast the variable. $\mathbf{w}_t$ is the $(p \times 1)$ vector of predetermined variables which we want to include in the model and $\mathbf{x}_t$ is the $(N \times 1)$ vector containing all the candidate variables that are potentially related to the dependent variable. $\boldsymbol{\beta}_w$ is a vector containing the coefficients for the predetermined variables and $\boldsymbol{\beta}_x$ is a vector containing the coefficients of the candidate variables. Moreover, ϵ_{t+h} is the error term for which we assume that the white noise property holds. The h-step ahead forecast of the dependent variable at time T is defined as $\hat{y}_{T+h|T} = \mathbf{w}_T'\hat{\boldsymbol{\beta}}_w + \mathbf{x}_T'\hat{\boldsymbol{\beta}}_x$, where $\hat{\boldsymbol{\beta}}_w$ is the estimated coefficient of the predetermined variables and $\hat{\boldsymbol{\beta}}_x$ the estimated coefficients of the candidate variables. Let $\mathbf{y} = (y_{1+h}, \ldots, y_{T+h})'$, $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_T)'$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)'$, then the model can be rewritten to matrix form as:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\beta}_w + \mathbf{X}\boldsymbol{\beta}_x + \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\epsilon}$ is the error term.

When the matrix $\mathbf{X}$ of candidate variables contains a substantial number of variables compared to the available observations, modeling the dependent variable as a linear combination of all candidate variables would entail estimating a significant number of parameters. Under the standard Ordinary Least Squares (OLS) assumptions of normal errors and strictly exogenous regressors, it can be derived that the mean squared forecast error will increase with the number of variables in the regression (Stock & Watson, 2006). This result shows that applying a standard OLS regression on Equation 2 will lead to forecasts with high variation for data with many candidate variables $\mathbf{x}_t$. However, we perform a standard OLS regression with all $\mathbf{w}_t$ and $\mathbf{x}_t$ as independent variables as a benchmark model, to which we will compare the performance of the other methods we apply. Previously, literature has shown that the use of a large cross-section of variables can improve traditional forecasting methods for macroeconomic data (Stock & Watson, 2002a; Artis et al., 2005; Smeekes & Wijler, 2018). Hence, factor modeling methods were developed and used in the last decades, such as by Stock & Watson (2002a), to be able to adopt the high dimensionality of the cross-section in macroeconomic data without losing forecasting power as in a standard OLS regression with high dimensionality (Stock & Watson, 2006). More recently, Smeekes & Wijler (2018) discussed the idea of using penalized regression techniques to forecast macroeconomic data with a high dimensionality, challenging the efficacy of traditional factor models. In the next sub-sections, a specific class of factor model and penalized regression model will be presented and discussed. These models were selected

in particular due to their strong empirical performance in the findings that Smeekes & Wijler (2018) presented. Particularly, these models seem to outperform other types of models in their respective categories.

4.2 Penalized regression estimators

4.2.1 Adaptive Lasso

Finding the correct variables for estimation and forecasting with high-dimensional data is a widely discussed subject. For complicated datasets with high collinearity, regularization is often used to directly estimate which variables need to be selected in the regression model (Tibshirani, 1996). In this paper, the Adaptive Lasso shrinkage estimator is utilized to obtain reliable model estimates. Zou (2006) specifies the Adaptive Lasso estimation model as an improved version of the original Lasso estimator. The authors derive the so-called oracle properties for the Adaptive Lasso, such as variable selection consistency and asymptotic normality of coefficient estimates. These properties imply that the method achieves performance as if having perfect knowledge of the true underlying model, even in complex and high-dimensional settings. This can also explain how the Adaptive Lasso resulted in the best performance in the paper of Smeekes & Wijler (2018). The Adaptive Lasso estimates the parameters according to the following objective function formulated by Zou (2006):

$$(\hat{\boldsymbol{\beta}}_w, \hat{\boldsymbol{\beta}}_x) = \arg \min_{(\boldsymbol{\beta}_w, \boldsymbol{\beta}_x)} \sum_{t=1}^{T} \left(y_{t+h} - \mathbf{w}_t' \boldsymbol{\beta}_w - \mathbf{x}_t' \boldsymbol{\beta}_x\right)^2 + \lambda_{\text{lasso}} \sum_{j=1}^{N} \frac{|\beta_{x,j}|}{\left|\hat{\beta}_{\text{Init},j}\right|}, \tag{4}$$

where λ is a penalty term that causes sparsity in the model, meaning that it tends to set some coefficients to zero. $\hat{\beta}_{\text{Init},j}$ is the initial estimator, which is determined using either OLS or ridge regression. Zou (2006) recommends using ridge regression over OLS when variables are highly correlated, as in practice using biased estimates for the weights in Adaptive Lasso will not matter much. Therefore, the ridge regression is used in this paper for computing the initial estimator $\hat{\boldsymbol{\beta}}_{\text{Init}}$. The ridge regression computes this estimator using the following objective function:

$$(\hat{\boldsymbol{\beta}}_w, \hat{\boldsymbol{\beta}}_{\text{Init}}) = \arg \min_{(\boldsymbol{\beta}_w, \boldsymbol{\beta}_{\text{Init}})} \sum_{t=1}^{T} \left(y_{t+h} - \mathbf{w}_t' \boldsymbol{\beta}_w - \mathbf{x}_t' \boldsymbol{\beta}_{\text{Init}}\right)^2 + \lambda_{\text{ridge}} \sum_{j=1}^{N} |\beta_{\text{Init},j}|^2 . \tag{5}$$

Furthermore, the Adaptive Lasso method performs subset selection, achieved by shrinking coefficient estimates towards zero. This allows the method to potentially be able to improve forecast performance by reducing the added variance of estimating parameters of irrelevant variables. Moreover, this method allows for model estimation even when the number of potentially relevant variables surpasses the number of observations, thus, when $N > T$. The Lasso model is also analyzed to compare its performance relative to the Adaptive Lasso. In a Lasso model, the coefficient $\hat{\beta}_{\text{Init},j}$ is equal to one, implying that the penalty applied to all coefficients is uniform. This indicates how applying weights to the penalty on each coefficient influences the performance of the lasso-type estimators.

4.2.2 Hyperparameter tuning

Zou (2006) discusses the method of cross-validation to apply the hyperparameter tuning for λ. The problem for regularization techniques such as the Adaptive Lasso is that the λ will be chosen to optimize the prediction error, but this does not necessarily result in the right model selection. Smeekes & Wijler (2018) use a simplified method where they construct a (100×1) vector of potential λ values and select λ which results in the best model fit according to the lowest MSFE. However, this does not necessarily result in the optimal model fit since the best λ might not be included in this vector. The authors of the original model specification derive that the Adaptive Lasso has oracle properties for choosing the correct model, thus, this might not be an issue relative to other regularization techniques. For simplicity, the method used by Smeekes & Wijler (2018) will be adopted in this research as the model performs better than some factor models without applying more advanced hyperparameter tuning methods.

4.3 Factor models

4.3.1 Specification

Factor models aim to summarize the candidate set into a reduced number of factors. The general goal of the factor specification is to express all candidate variables as a linear function of a small number of factors. By constructing this relationship, all the information of the candidate variables can still be rendered, while reducing the dimensionality by only using the factors in the forecasting of the dependent variable. The use of only a restricted amount of factors simplifies the interpretation in the estimation of the dependent variable. The specification of the factor model for the variables in the candidate set can be expressed as:

$$\mathbf{x}_t = \gamma(L)\mathbf{f}_t + \mathbf{e}_t, \tag{6}$$

where $\mathbf{\Gamma}(L) = (\gamma_1(L), \ldots, \gamma_N(L))'$, and $\gamma_i(L) = (\gamma_{i,1}(L), \ldots, \gamma_{i,s}(L))'$ with $\gamma_{i,j}(L)$ being a lag polynomial describing how variable i influences dynamic factor j (Artis et al., 2005; Smeekes & Wijler, 2018). Moreover, $\mathbf{f}_t$ is the vector of common factors and $\mathbf{e}_t$ is the vector of error terms. This model is commonly referred to as the dynamic factor model and Forni et al. (2000) have shown the possibility of estimation factors through PCA with the lag polynomial being infinite.

Previous literature in the domain of factor forecasting assumes the lag polynomial to be finite (Artis et al., 2005; Smeekes & Wijler, 2018), which we also assume in our model specification. As the lag polynomials $\gamma_{i,j}(L)$ are finite, the model can be written in the following static form:

$$\mathbf{x}_t = \mathbf{\Gamma}\mathbf{F}_t + \mathbf{e}_t, \tag{7}$$

where $\mathbf{\Gamma}$ is a vector that contains the coefficient in $\mathbf{\Gamma}(L)$ and $\mathbf{F} = (\mathbf{f}'_t, ..., \mathbf{f}'_{t-q})'$. In this static specification, the lag polynomials $\gamma_i(L)$ have order at most q. Stock & Watson (2002b) show that the factor model can be estimated consistently through the Principal Components method of X. The general equation of the estimation of $\mathbf{F}$ and $\mathbf{\Gamma}$ through the PC model is given by:

$$(\hat{\mathbf{\Gamma}}^k, \hat{\mathbf{F}}^k) = \arg\min_{\mathbf{\Gamma}^k, \mathbf{F}^k} \sum_t (\mathbf{x}_t - \mathbf{\Gamma}^k \mathbf{F}_t^k)' \mathbf{\Omega}^{-1} (\mathbf{x}_t - \mathbf{\Gamma}^k \mathbf{F}_t^k), \tag{8}$$

subject to the conditions:

- $\mathbf{\Gamma}^{k\prime}\mathbf{\Gamma}^{k\prime}/N = \mathbf{I}_k$
- $\mathbf{F}^{k\prime}\mathbf{F}^{k\prime}$ diagonal
- $\mathbf{\Omega} = I_N$.

If these conditions are satisfied, the PC model can identify a $(T \times k)$ matrix of k estimated factors and $(N \times k)$ matrix of estimated factor loadings (Stock & Watson, 2002a). The estimation method of the PC model will also be adopted to analyze its relative performance with the SPC model.

In order to understand how the factor model is used to forecast dependent variable y_{t+h}, it is necessary to substitute the general factor model in Equation 7 into the macroeconomic specification in Equation 2 as shown in Equation 9 (Smeekes & Wijler, 2018), which is given by:

$$\begin{aligned} y_{t+h} &= \mathbf{w}'_t\boldsymbol{\beta}_w + \mathbf{x}'_t\boldsymbol{\beta}_x + \epsilon_{t+h} \\ &= \mathbf{w}'_t\boldsymbol{\beta}_w + \mathbf{F}'_t\mathbf{\Gamma}'\boldsymbol{\beta}_x + \mathbf{e}_t\boldsymbol{\beta}_x + \epsilon_{t+h} \\ &= \mathbf{w}'_t\boldsymbol{\beta}_w + \mathbf{F}'_t\boldsymbol{\beta}_f + u_{t+h}. \end{aligned} \tag{9}$$

Using this specification of the macroeconomic model, a h-step ahead forecast can be constructed using a two-step approach. Firstly, the factor $\hat{\mathbf{F}}_k$ and factor loading $\hat{\mathbf{\Gamma}}_k$ are estimated. Subsequently, they are used to construct the forecast $\hat{y}_{t+h|T} = \mathbf{w}'_t\hat{\boldsymbol{\beta}}_w + \hat{\mathbf{F}}'_t\hat{\boldsymbol{\beta}}_f$. This second estimation can be performed using OLS, as the dimensionality is lower given that the factors are used. Applying this two-step approach, the factor model specification will be used to construct forecasts for macroeconomic series with large amounts of variables.

4.3.2 Sparse Principal Component estimation

The classical method for estimating factors is the standard PC estimation as shown in Equation 8. However, alternative methods and modifications are used to estimate the factors in order to achieve improved prediction accuracy. Boivin & Ng (2006) show in their analysis that the PC model is sensitive to the idiosyncratic component of the errors and higher cross-correlation leads to less efficient factor forecasts. This is highly likely within a macroeconomic framework characterized by variables that can exhibit significant variability due to unforeseen effects, therefore, it should be taken into account when deciding on the best forecasting methodology. The article also questions whether it is always advantageous to incorporate all candidate variables $\mathbf{x}_t$. Moreover, Zou et al. (2006) and Smeekes & Wijler (2018) further criticize the standard PC model as the factors $\mathbf{F}$ are made of linear combinations of all the candidate variables $\mathbf{x}_t$, which makes the factors difficult to interpret. Therefore, using the SPC model because of its forecasting accuracy seems to be consistent with the empirical findings of Smeekes & Wijler (2018).

Zou et al. (2006) formulated the Sparse Principal Component equation, where the ordinary Principal Component estimation is modified and combined with the lasso-type regression. It is important to note that the factor model that produces the best forecasting accuracy also adopts penalized regression in some form (Smeekes & Wijler, 2018). In this case, the SPC model tries

to estimate the factors using a sparse amount of candidate variables instead of using the whole set of variables. The formulation for the Sparse Principal Component estimation is given by Zou et al. (2006) and is reformulated as follows in the context of the factor model:

$$(\hat{\mathbf{\Gamma}}^k, \hat{\mathbf{F}}^k) = \arg\min_{\mathbf{\Gamma}^k, \mathbf{F}^k} \sum_t (\mathbf{x}_t - \mathbf{\Gamma}^k \mathbf{F}_t^k)' \mathbf{\Omega}^{-1} (\mathbf{x}_t - \mathbf{\Gamma}^k \mathbf{F}_t^k) + \delta \sum_{j=1}^{k} \|\mathbf{\Gamma}_j^k\|^2 + \sum_{j=1}^{k} \delta_{1,j} \|\mathbf{\Gamma}_j^k\|_1, \qquad (10)$$

where the same δ is used for all k factors, while distinct $\delta_{1,j}$ values are permitted to penalize the loadings of individual factors. Zou et al. (2006) note that PC estimation is a type of regression model, hence, penalized regression models can be implemented such as in Equation 10. Using Equation 10, the estimation directly computes which variables are relevant for the factors and Zou et al. (2006) shows that this method correctly identifies the important variables. Hence, this gives a strong tool to estimate the factors, keeping interpretability high due to dimension reduction and potentially achieving better forecast accuracy.

In order to solve Equation 10, the R package *sparsepca* is used, which provides the function SPCA. This function includes α and β which represent the sparsity controlling parameter and the amount of ridge shrinkage to apply, respectively. A high value of α will lead to a sparser component and a high value of β will imply more ridge shrinkage. These parameters are both tuned, firstly using a (100×1) vector of potential α in which the α is selected which minimizes the squared error. When the right α is determined, a (100×1) vector of potential β values is used to determine the best performing β in a similar manner.

4.3.3 Determining the number of factors

Traditionally, the factor model is used frequently for forecasting macroeconomic data as discussed more thoroughly in Section 2. In a variety of research undertaken in this area, the number of factors was often taken as given. This is not necessarily consistent over different types of datasets used in a macroeconomic context. Hence, it would be more useful to determine the number of factors coherent with the data that is used in the analysis. Bai & Ng (2002) elaborate on this and construct a method to select the number of factors for Equation 7. They note that the traditional method of using AIC and BIC as information criteria is not consistent for datasets that have a large T and N, as both selection criteria are only functions of one of them respectively. Consequently, the authors try to specify novel criteria to select the correct number of factors in the model for large values of T and N. This is specifically relevant for our research as the goal is to forecast within the framework of macroeconomic data with a large number of variables. Bai & Ng (2002) discuss several constructed selection criteria to acquire the correct number of factors. In the analysis, they allow for some cross-correlations in the errors and conclude that these criteria select the number of factors approximately. One of the issues discussed about the method is that it primarily works for static factor models and, therefore, Alessi et al. (2010) provided a better specification of the selection criteria. They demonstrate that the newly constructed selection criterion, which includes an extra tuning parameter, is more effective in practical contexts with large idiosyncratic components and is particularly advantageous for our research. Alessi et al. (2010) specify the selection criterion for the number of

factors which we will use as follows:

$$IC_{c,n}^{T}(k) = \log[V(k)] + ck(\frac{n+T}{nT})\log(\frac{nT}{n+T}), \quad c \in \mathbb{R}^{+} \tag{11}$$

where $V(k)$ is the optimization problem of the factor and factor loadings as described in Equation 8. This selection criterion adds the penalty term which is described by the second element on the right-hand side of the equation. The penalty term envelops the number of variables, time periods, and factors chosen. Furthermore, the extra tuning parameter c is added to the penalty term to give better estimations for the number of factors. Ultimately, the number of factors is chosen according to the following optimization problem:

$$\hat{r}_{c,n}^{T} = \underset{0 \leq k \leq r_{max}}{\arg\min}\ IC_{c,n}^{T}(k), \tag{12}$$

where $\hat{r}_{c,n}^{T}$ is the number of factors chosen for the forecasting model. A small value for the parameter c leads to an overestimation of the number of factors $\hat{r}_{c,n}^{T}$ as the penalty is relatively low, and vice versa. The best-performing hyperparameter value is selected by constructing a (20×1) vector and choosing the value that results in the lowest selection criterion.

5 Results

We forecast two macroeconomic variables on datasets to investigate the efficacy of machine learning techniques relative to factor models. We use datasets from Brazil and Costa Rica, respectively, which have been appropriately preprocessed as explained in Section 3. The code was implemented in R, using the packages *glmnet* and *sparsepca*. Particularly, the *sparsepca* package uses function SPCA to solve Equation 10. This function includes α and β which represent the sparsity controlling parameter and the amount of ridge shrinkage to apply respectively. A high value of α will lead to sparser component and a high value of β will imply more ridge shrinkage. These parameters are both tuned, firstly using a (100×1) vector of potential α in which the α is selected which minimizes the squared error. When the right α is determined, a (100×1) vector of potential β values is used to determine the best performing β in a similar manner.

5.1 Forecast Results

This section discusses the results of the forecasting performance for the penalized regression models and traditional factor models. Using a 10-year rolling window, one-step ahead forecasts are constructed for the quarterly data from Brazil and Costa Rica, respectively, for the dependent variables $RGDP$ and CPI. Importantly, it seems that, in line with findings from (Smeekes & Wijler, 2018), there are instances in which penalized regression models are able to outperform traditional factor models, such as for CPI in Brazil. However, in most cases, factor models remain competitive.

In Table 1, the forecasting efficacy of lasso-type models and Principal Component type models are examined using the Relative Mean Square Forecast Error (ReMSFE), relative to the benchmark OLS model. The Adaptive Lasso outperforms all other models in terms of ReMSFE for the CPI dependent variable of Brazil. Moreover, it can be seen that the ReMSFE of the

Table 1: Relative Mean Squared Forecast Error (ReMSFE) for Brazil and Costa Rica.

	Brazil					Costa Rica				
	OLS	Lasso	AdaLasso	PC	SPC	OLS	Lasso	AdaLasso	PC	SPC
$RGDP$	1	1.40	0.775	0.707	0.707	1	1.11	1.16	0.820	0.820
CPI	1	0.622	0.633	0.640	0.640	1	0.625	0.614	0.567	0.567

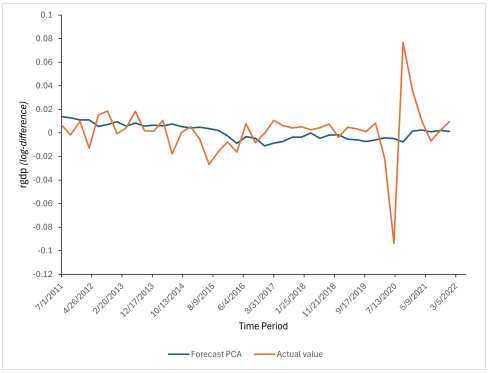
Note: Numerical entries in this table are obtained relative to the benchmark OLS method.

factor models using PC and SPC is the lowest for the dependent variable $RGDP$. This partly aligns with the findings from (Smeekes & Wijler, 2018), reinforcing the argument that Adaptive Lasso is particularly effective for nominal series such as prices indices as CPI.

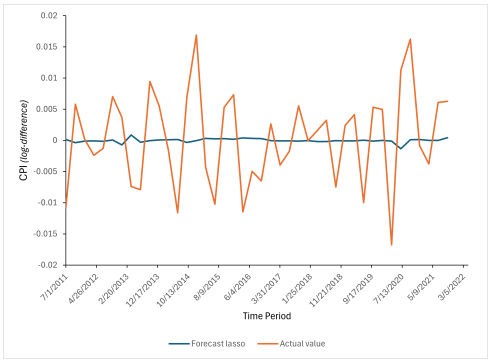
Conversely, in Costa Rica, PC and SPC models exhibit superior performance over the lasso-type models for both dependent variables, which suggests that factor models might be more adept at capturing the dynamics influencing GDP and CPI. An explanation for this phenomenon could be the smaller size of Costa Rica relative to both Brazil and the US. Hence, only using independent variables that are economic indicators might not capture the global fluctuations, i.e., exogenous shocks, in the economy. A factor model representation such as PC or SPC seems to capture such fluctuations more effectively and results in a better forecasting performance than penalized regression models. Table 1 shows that both lasso-type estimators seem to perform poorly relative to the OLS benchmark for $RGDP$. This is in line with results from Smeekes & Wijler (2018), who found that the performance of penalized regression models is predominantly superior for price indices; factor models generally outperform other models for real series. For $RGDP$, the Adaptive Lasso and Lasso models perform similarly, with ReMSFE values of 1.11 and 1.16, respectively, suggesting a marginal advantage for the Adaptive Lasso. Both the PC and SPC models yield identical ReMSFE values of 0.820 for $RGDP$, indicating that, as with Brazil, the addition of sparsity does not improve forecasting accuracy for $RGDP$ in Costa Rica. These findings are similar to those for forecasting CPI, where ReMSFE yields 0.567 for both the PC and SPC models. However, in line with the finding from Smeekes & Wijler (2018) that penalized regression methods perform better for price indices, the performance of the Adaptive Lasso and Lasso models are closer to that of the PC and SPC models than for $RGDP$.

Notable from Table 1 is the parallel performance of PC and SPC across both countries and indicators, as mentioned above. This result suggests that, in the context of forecasting major economic indicators in Brazil and Costa Rica, the additional complexity of imposing sparsity does not necessarily translate into improved forecasting accuracy. The main variables extracted by PCA could already be the most significant predictors, thus, sparsity might exclude variables that contribute meaningfully to the prediction, therefore not improving the forecasting performance. This outcome underscores the importance of identifying underlying economic structures to suitably match the choice of methodology to the data.

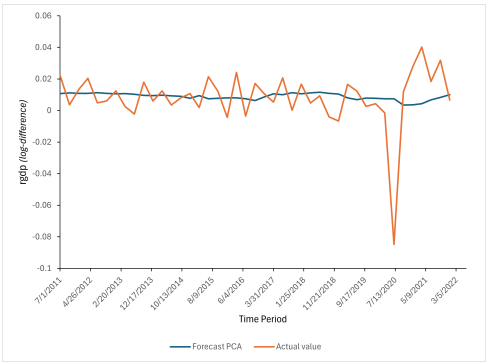
In Figure 1, the one-step forecasts are plotted against the actual values for the best-performing model of each dependent variable. The graphed forecasts for all other models can be found in Appendix **??**. Figure 1 shows that the one-step ahead forecasts for the quarterly data do not adequately capture the business cycles in the dependent variables. The actual value of the dependent variables fluctuates significantly more than the forecasted values for all graphs, which



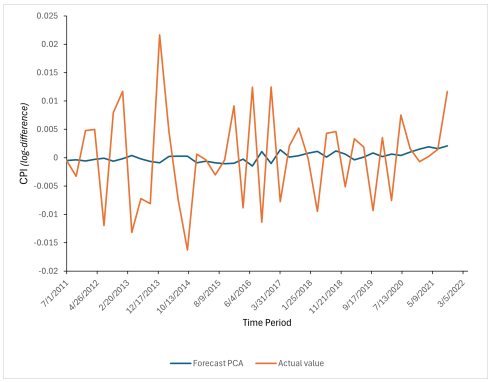
(a) One-step forecast of PCA for $RGDP$ in Brazil.



(b) One-step forecast of Lasso for CPI in Brazil.



(c) One-step forecast of PCA for $RGDP$ in Costa Rica.



(d) One-step forecast of PCA for CPI in Costa Rica.

Figure 1: One-step ahead forecast values against actual realized values for the best-performing models.

can be seen by the close-to-constant lines for the forecasts. This could be explained by the presence of sparsity in the data, where only a few variables explain the variance in the response variable. When the data is sparse, the dimensionality reduction selects too few variables for the models to be able to accurately prdict the fluctuations in the dependent variable. This can be seen both in the penalized regression-type models and PC-type models. The penalized regression models only select a sparse number of variables due to the lasso regularization, and the factor models summarize the data in a few numbers of factors. Hence, it can be concluded that the sparsity of the data for developing countries seems to hinder the model from predicting the business fluctuations accurately. Another reason for the poorly performing forecasts could be that there are too few observations in the datasets for the model to train correctly. Therefore, the models lack sufficient data to facilitate accurate learning and subsequent predictions. Regression models and more specifically machine learning models generally perform better in forecasting when more training data is provided. Consequently, it seems that in the context of scarce data for developing economies, the models seem to underperform. The most frequent data which could be found for these economies are quarterly data starting in 2001, both for Brazil and Costa Rica. It appears that the limited number of observations is insufficient for accurate forecasting. Furthermore, there may be a deficiency in the specified forecasting model, leading to inaccurate predictions of the dependent variables. More research should be performed

to understand which independent variables and lags can be best incorporated in the forecasting model. Lastly, a key explanation to these results and also the hyperparameter tuning results, see Section 5.2 and 5.3, is the fact that data of developing economies is noisy. The idiosyncratic components might be large and serially-correlated, which makes it challenging for the models to forecast accurately. This noise is also results from the economies smaller role in the global economy, and therefore, being more susceptible to exogenous shocks. This could be related to the model-specification problems discussed previously, since incorporating independent variables from other significant economies, such as the US, might explain this noise.

Additionally, it is important to note that the data transformation of the variables might not have been thorough enough. Smeekes & Wijler (2018) transform all variables independently with regards to what the variable truly represents. In our research, the independent variables have been normalized and the dependent variables have been transformed by log-differences. Thus, further research is necessary to understand how the variables can best be transformed for the models to be able to forecast the business cycles of the dependent variables more accurately.

However, while the result in Figure 1 shows the relatively poor forecasting performance of the model, it does not mean that the comparative analysis of the models is less pertinent. Consistent independent variables and transformations are utilized across all models and their respective forecasts. Even though the forecasts of the models might not be completely accurate, the same data and techniques are used to perform the one-step ahead forecasts of the dependent variables. Hence, the relative results from Table 1 are still relevant and conclusions can still be drawn on the comparative performance of penalized regression and factor models.

5.2 Hyperparameter tuning of lasso-type models

Regarding the estimates for the Lasso and Adaptive Lasso models, we apply the method used by Smeekes & Wijler (2018), implementing time series cross-validation to determine λ in Equations 4 and 5. For both the Lasso and Adaptive Lasso regression, the selection of the tuning parameters is essential for the performance of the model. Identifying the appropriate hyperparameter is crucial because penalized models rely on this parameter to select the correct independent variables used for forecasting. If the correct hyperparameter is not identified, the lasso-type models can select the incorrect variables, which will result in the inaccurate forecasts.

For the Lasso regression, 100 potential λ candidates are generated, spanning from 0.001 to 100. We employ a time series cross-validation scheme to find the optimal λ, where each dataset is divided into a training set and a validation set, containing approximately two-thirds and one-third of the observations, respectively. Each λ candidate in the grid is evaluated by fitting the Lasso regression and calculating the mean squared forecast error. The λ that yields the lowest MSFE is selected as the optimal parameter. Specifically, for the economic indicators $RGDP$ and CPI in Brazil, the optimal λ values were found to be 0.0023 and 0.0025, respectively. In contrast, for Costa Rica, the optimal λ values for $RGDP$ and CPI were determined to be 0.001 and 0.0013. Figure 2 illustrates the relationship between the MSFE and different λ values. The y-axis represents the mean square forecast error, while the x-axis displays various λ values. Overall, the performance of candidate λ values appears to be consistently stable, with the exception of instances where λ is close to zero. The minimum values, as previously

mentioned, are therefore the optimal λ selections.

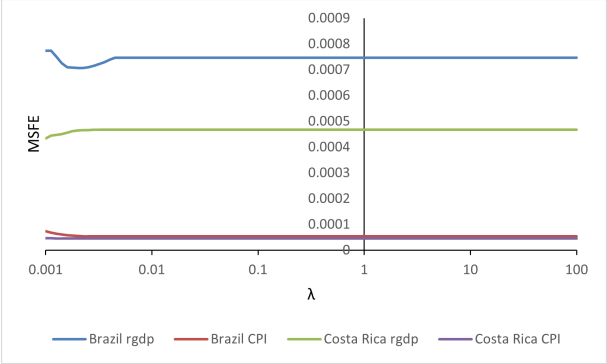
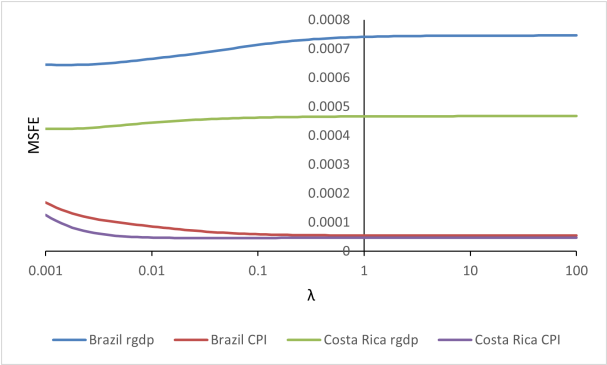


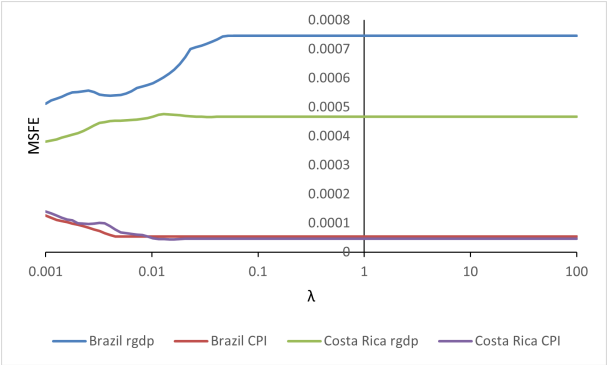
Figure 2: Hyperparameter tuning for the standard Lasso model.

The tuning of the Adaptive Lasso involves two steps: first, selecting an initial λ for the ridge regression, and second, determining the λ for the Adaptive Lasso regression. The first step involves running a ridge regression to obtain an initial estimate of the coefficients. This is done using the same time-series cross-validation framework as described above. This λ is selected such that the MSFE is minimized in the ridge regression. The obtained coefficients are then used as the weights for the Adaptive Lasso. The optimal ridge λ values according to the time series cross-validation are 0.0018 and 100 for $RGDP$ and CPI in Brazil, respectively. In Costa Rica these values are 0.0014 and 0.0368, respectively. Figure 3a depicts the relationship between the MSFE and the λ values. Notably, the λ value that minimizes the MSFE for CPI in Brazil stands out as larger compared to other λ values. This implies that for CPI a significant amount of regularization is applied to the regression. The substantial level of regularization applied indicates a strong bias towards simpler models with smaller coefficients. This level of regularization heavily penalizes large coefficients, effectively constraining the complexity of the model and reducing its tendency to overfit the training data.

In the subsequent step, the coefficients derived from the ridge regression, which minimizes its MSFE, are incorporated as weights into the Adaptive Lasso regression. The time-series cross-validation technique is utilized to ascertain the λ value for the Adaptive Lasso regression that minimizes the MSFE for each economic indicator, both for Brazil and Costa Rica. Specifically, for $RGDP$ and CPI in Brazil, the λ values for the Adaptive Lasso regression are determined as 0.001 and 0.0051, respectively. Conversely, for the same indicators in Costa Rica, the respective λ values are identified as 0.001 and 0.0145. Figure 3b illustrates the relationship between the MSFE and the Adaptive Lasso λ values. Notably, the minimum λ values for $RGDP$ in both countries reside at the lower bound of the domain, suggesting the potential existence of an optimal λ value even smaller than 0.001. Conversely, for CPI, these minima are slightly larger in value. These results show that the effectiveness of the Adaptive Lasso method in selecting the appropriate λ value is influenced by the characteristics of the data and the context in which it is applied. This deduction stems from the argument that the variability in the λ parameter selection is solely dependent on the type of data employed, which varies in the selected variables, the preprocessing methods applied, and the economic context of the respective country.



(a) Tuning hyperparameters ridge regression.



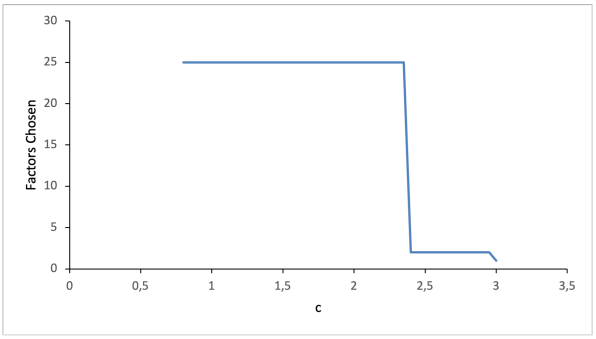
(b) Tuning hyperparameters Adaptive Lasso regression.

Figure 3: Tuning hyperparameters for Adaptive Lasso forecasting.

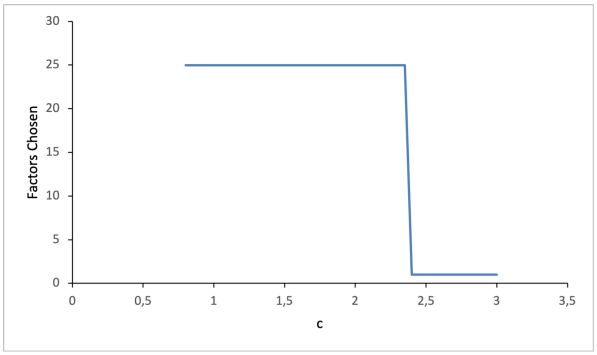
5.3 Factor model tuning

The number of factors included in the Principal Component type models are chosen based on the selection criteria of Alessi et al. (2010). The value of this criteria c, which indicates the penalty term on the number of factors, is tuned in order to give a better estimation of the number of factors. The result of this tuning can be seen in Figure 5. When c has a small value it will lead to an overestimation of the number of factors as the penalty term is low. We observe that for Brazil the factors chosen decrease from 25 to 2 when c reaches a value of 2.4. Moreover, for Costa Rica, we find that the number of factors chosen decreases from 25 to 1 when c reaches 2.4. Thus, for Brazil, both variables $RGDP$ and CPI. the model selects two factor, and for Costa Rica, only one factor is selected for both variables. These values are chosen by the elbow method, illustrated in Figure 4. Choosing the number of factors in Equation 12, r_{max} was set to 25. For low penalization of the number of factors, it can indeed be seen that the maximum amount of factors is chosen. Only when the penalization tuning parameter c is set to high values, there is a significant reduction in the number of factors selected. This is not necessarily what would be expected as the goal of PC and SPC models is to summarize the independent variables in a sparse number number of variables. Figure 4 seems to show that summarizing the data in a few factors does not result in better performance as 25 factors are chosen initially. One of the explanations behind this is the fact that the dataset could be noisy resulting in large idiosyncratic components. This could then lead to the selection criterion not performing well in selecting the number of factors for the forecast. Moreover, this could also be due to the fact

that the data transformation was not tackled thoroughly enough. The reason for the different number of factors being selected for the two countries may be caused by the complexity and size of the economy of Brazil compared to that of Costa Rica. This suggests that variations in economic structure impact the manner in which dependent variables are explained by factors, consequently influencing the application of econometric models.

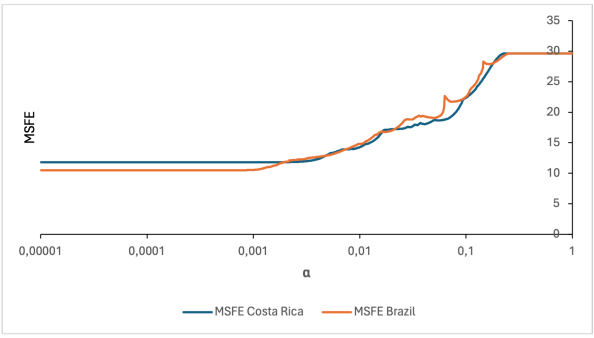


(a) Tuning of number of factors used in PC/SPC for Brazil.

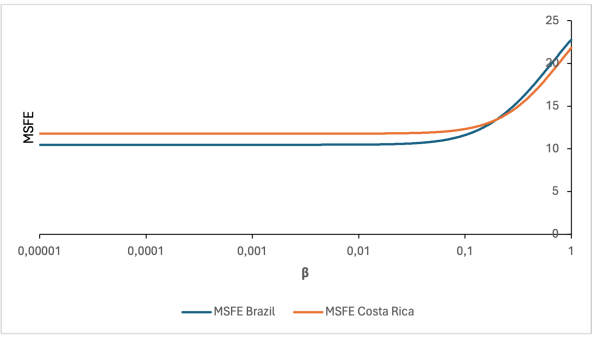


(b) Tuning of number of factors used in PC/SPC for Costa Rica.

Figure 4: Tuning of number of factors used PC/SPC for Brazil and Costa Rica.



(a) Tuning of alpha in SPC.



(b) Tuning of beta in SPC.

Figure 5: Tuning of alpha and beta in SPC.

Figure 5 shows the squared error of the estimation fit of Equation 7 for the different values of α and β. The values α and β are the tuning parameters for the estimation of the Sparse Principal Component model. For both α and β the MSFE is constant for small values of α and β, but increases when their respective values become bigger. α indicates the amount of sparsity, meaning that higher values lead to sparser components. Moreover, β decides the amount of ridge shrinkage. Figure 5 indicates that the α and β will not have large values since this causes too much sparsity and shrinkage. Too much shrinkage will cause variables to be excluded that are important for predicting the outcome, leading to the model missing information. This additionally reinforces the general forecasting results that show no difference between PC and SPC models in the application.

6 Conclusion and Discussion

This paper explores the performance of lasso-type models and Principal Component-type models for forecasting macroeconomic variables in developing economies, specifically in Brazil and Costa Rica. Through a comparative analysis of these forecasting methodologies, methods for macroeconomic forecasting are discussed in the economic context of developing economies with volatile data and limited data availability. This research replicates the methods of Smeekes & Wijler (2018) in the setting of developing economies, to ascertain whether the findings can be extended beyond the US. It is discussed in this paper that when choosing methods to forecast economic variables, it is important to consider the specific needs and challenges of developing countries. By exploring the strengths and weaknesses of these methods, this study helps improve how economic forecasting is approached in countries where capturing economic trends is important but challenging due to the scarce data availability.

For Brazil, the results show that the Adaptive Lasso method performs better in terms of forecasting the Consumer Price Index. Furthermore, for Real Gross Domestic Product, the PC and SPC models perform best but with no large performance difference relative to the Adaptive Lasso. This suggests that in larger economies, where the potential for variable relevance and sparsity in relationships exists, penalized regression methods like Adaptive Lasso can provide an advantage by effectively selecting the most relevant predictors. On the other hand, in Costa Rica, factor models like PC and SPC exhibited superior performance and surpassed the lasso-type models. This result shows that factor models might work well in smaller economies, which could be caused by their ability to simplify complex data into a few factors that show the main economic movements. The forecasting results also demonstrate that the Adaptive Lasso performs better when it comes to the prediction of price indices which is in accord with the analysis of Smeekes & Wijler (2018). Moreover, it was found that for both countries the PC and SPC models had similar performance indicating that adding sparsity does not lead to better forecasts of every macroeconomic variable in the context of developing economies. This is in accordance expectation when using restricted data with fewer variables and observations. Generally, the literature suggests that the SPC models are an improvement of PC models when the dimensionality in the data is high. Consequently, policy makers should carefully evaluate the methods they use for forecasting economic variables based on the availability and structure of the data. In our research, it seems the standard PC model is preferable over the SPC model, both in Brazil and Costa Rica.

While providing insights into the applicability of Adaptive Lasso and the Sparse Principal Component model in forecasting macroeconomic variables in developing economies, this paper also faces certain limitations. In Section 5.1, we show that the forecasts of the models are not able to predict the general business cycle of the macroeconomic indicators accurately. We believe that three primary factors contribute to the inadequate forecasting of the built models: the limited steps undertaken for data preprocessing, the noise in the data, and the scarcity of the data. As discussed in Section 5, the transformation on the independent variables could have been extended beyond the Yeo-Johnson transformation for the independent varaibles to improve forecasts. These data preprocessing techniques also affect the chosen hyperparameters, thus, it is one of the main limitations to this research. It is recommended to execute further research

on optimal data transformation techniques for accurate forecasting. Additionally, the research is limited in the fact that the results are easily manipulated by choosing different values and methods for the hyperparameter tuning used for both SPC and Adaptive Lasso. The belief is that this is caused by the data being too noisy, resulting in less reliable forecasts. The volatile nature of data in developing countries poses significant challenges for models to discern which variables affect the fluctuations, ultimately leading to poor forecasts. Furthermore, a limitation in this paper is the limited availability of data in developing countries, both in the number of variables available and the frequency of data. We conclude that both models provide exhibit poor forecasting performance under scarce data. It seems that reducing the dimensionality hinders the ability of the models to forecast the business cycles appropriately. This limitation complicates the application of the forecasting methods and also affects the robustness and depth of the analysis. It is desirable to have a large number of observations since the model can be trained more precisely as a result. Therefore, the primary policy recommendation is that these models are not suitable for forecasting data from developing economies due to the high amount of noise and scarcity of observations and variables in the data. This is opposed to the results of Smeekes & Wijler (2018), where they give a strong conclusion that penalized regression models are able to achieve strong forecasting performance for certain series. In our paper it is clear that in the context of developing economies, given the limited data, the models should not be used with the purpose of accurate forecasting.

Lastly, the research focuses on two specific types of models, the lasso-type models and two PCA-type models, based on the findings of Smeekes & Wijler (2018). While these models were chosen due to their effectiveness in previous studies, this focus may limit the extent to which the findings can be generalized. Other forecasting models or methodologies might offer different insights or perform differently under the same conditions, which we suggest for further research. Further research could also focus on how these forecasting methods work in many other developing and developed countries. This gives more insight into the usefulness of these methods everywhere or if they need changes to fit the unique economic situation of different countries.

References

Alessi, L., Barigozzi, M. & Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters*, *80*(23–24), 1806–1813.

Artis, M. J., Banerjee, A. & Marcellino, M. (2005). Factor forecasts for the uk. *Journal of Forecasting*, *24*(4), 279–298.

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

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

Chuku, C., Simpasa, A. & Oduor, J. (2019). Intelligent forecasting of economic growth for developing economies. *International Economics*, *159*, 74–93.

Croushore, D. (2006). Forecasting with real-time macroeconomic data. *Handbook of economic forecasting*, *1*, 961–982.

Fildes, R. & Stekler, H. (2002). The state of macroeconomic forecasting. *Journal of macroeconomics*, *24*(4), 435–468.

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

Granger, C. W. J. & Newbold, P. (1977). *Forecasting economic time series*. Academic press.

Hoffmaister, A. W. & Roldos, J. E. (2001). The sources of macroeconomic fluctuations in developing countries: Brazil and korea. *Journal of Macroeconomics*, *23*(2), 213–239.

Huang, J., Ma, S. & Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 1603–1618.

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

Lawley, D. N. & Maxwell, A. E. (1973). Regression and factor analysis. *Biometrika*, *60*(2), 331.

Leon-Gonzalez, R. & Thu, L. H. (2021). Forecasting macroeconomic variables in emerging economies. *Journal of Asian Economics*, *77*, 101403.

Medeiros, M. C. & Mendes, E. F. (2016). 1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics*, *191*(1), 255–271.

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. (1999). Business cycle fluctuations in us macroeconomic time series. *Handbook of macroeconomics*, *1*, 3–64.

Stock, J. H. & Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*(460), 1167–1179.

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

Stock, J. H. & Watson, M. W. (2006). Chapter 10 forecasting with many predictors. In *Handbook of economic forecasting* (pp. 515–554).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Yeo, I.-K. & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

Zou, H., Hastie, T. & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265–286.

Appendix

Table 2: Description of variables

Variable Name	Description
Dependent	
RGDP	Real Gross Domestic Product (national currency)
CPI	Consumer Price Index
Independent	
GDP	Gross domestic product (in US dollars, millions)
GFCF	Gross fixed capital formation (in US dollars, millions)
EXP	Exports of goods and services (in US dollars, millions)
IMP	Imports of goods and services (in US dollars, millions)
NPIRW	Net primary income from the rest of the world (national currency, millions)
GNI	Gross national income (national currency, millions)
NCuTRW	Net current transfers from the rest of the world (national currency, millions)
NCaTRW	Net capital transfers from the rest of the world (national currency, millions)
GCF	Gross capital formation (national currency, millions)
NLNB	Net lending / net borrowing (national currency, millions)
PFCE	Private Final Consumption Expenditure (national currency)
GFCE	Government Final Consumption Expenditure (national currency, millions)
M1	M1 money stock (national currency)
M3	M3 money stock (national currency)
MANU	Total manufacturing (in US dollars, millions)
CONSTR	Total construction (in US dollars, millions)
AvgBRRUSD	Average BRR/US exchange rate
AvgCRCUSD	Average CRC/US exchange rate
NAFA	Net acquisition of financial assets (national currency)
OINAFA	Other investment net acquisition of financial assets (national currency)
OINIL	Other Investment Net Incurrence of Liabilities (national currency)
FAN	Financial account net (national currency)
TDServices	Total debt in services (national currency)
TDGoods	Total debt in goods (national currency)
CTB	Capital Transfer Balance (national currency)
CAB	Capital Account Balance (national currency)
OIDSOneYear	Amount Outstanding Due within One Year of International Debt Securities for All Issuers, Residence of Issuer in specified country
OIDSGGAM	Amount Outstanding of International Debt Securities for Issuers in General Government Sector, All Maturities, Residence of Issuer in specified country
CMIR	Call Money/Interbank rate
GPEnergy	Global Price of Energy (period averages in nominal US dollars)
SWUI	Smoothed World Uncertainty Index

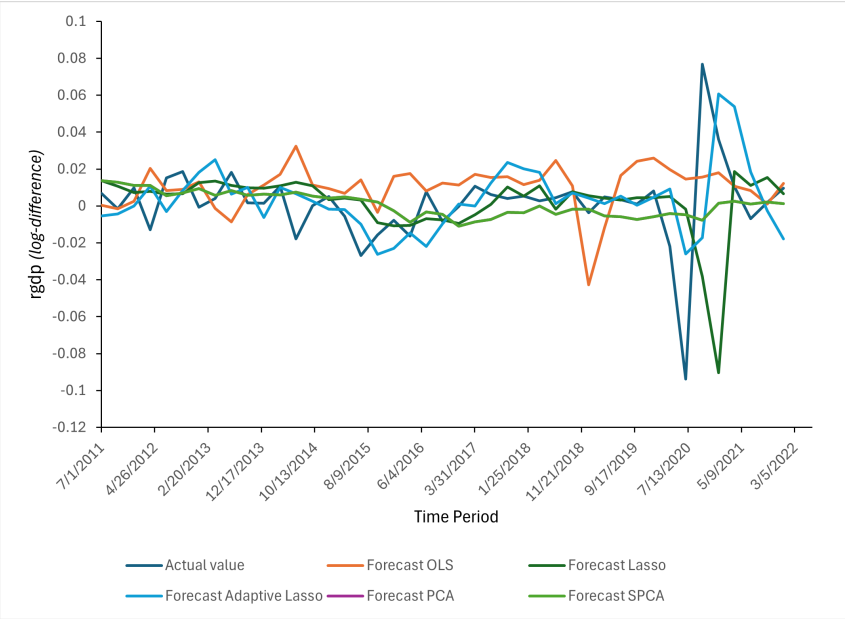


Figure 6: One-step forecast of all models for $RGDP$ in Brazil.

Figure 7: One-step forecast of all models for CPI in Brazil.

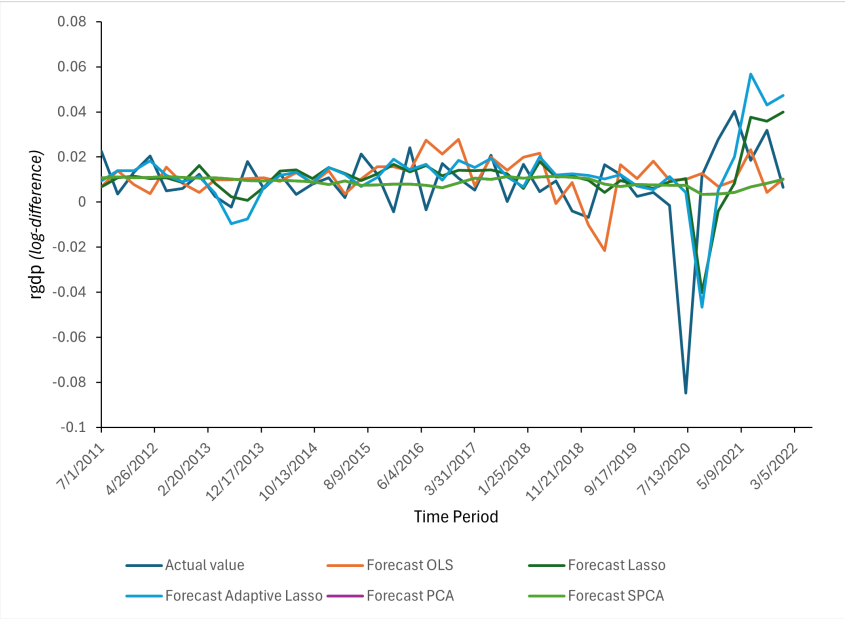


Figure 8: One-step forecast of all models for $RGDP$ in Costa Rica.

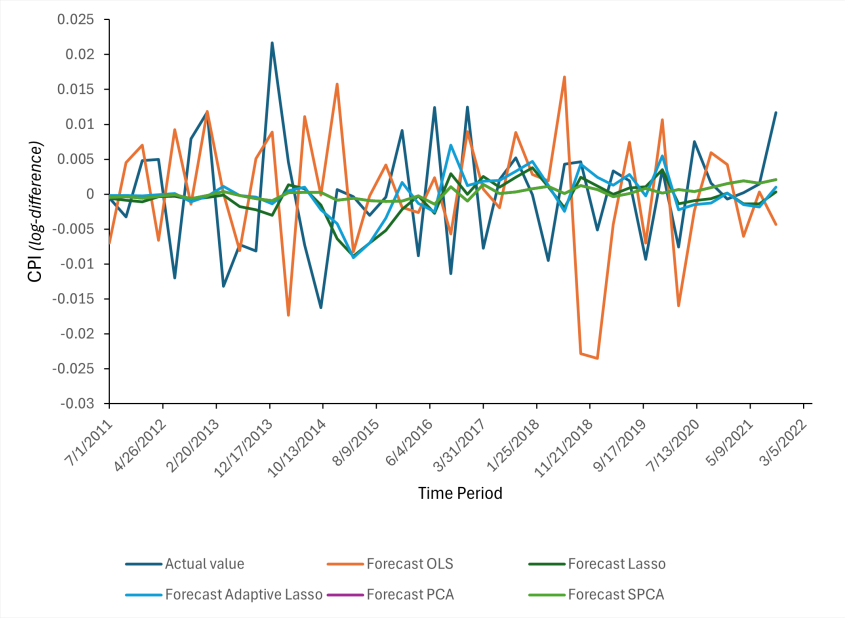


Figure 9: One-step forecast of all models for CPI in Costa Rica.