Erasmus University Rotterdam

Erasmus School of Economics

Bachelor's Thesis BSc^2 Econometrics and Economics

Impact of dimensionality reduction and factor selection methodology on African GDP forecasts

Maximilian Schnabl (467055)

Supervisor: Dr. Philip Hans Franses

Second Assesor: Dr. Carlo Cavicchia

June 29, 2021

Abstract

Constructing accurate forecasts from economic data is essential for economic policy decision-making. But low-frequency and high-dimensionality of such data can pose a serious challenge for classical statistical techniques. Kim & Swanson (2018) give an extensive overview of the most promising methodologies for analyzing such data. In this paper, a dataset containing yearly GDP growth rates of 52 African countries is used to replicate the findings of Kim & Swanson (2018) and answer how factor selection and dimensionality reduction techniques impact forecasting accuracy in terms of $MSFE$. In line with prior literature, it is found that factor-based models outperform AR-type models. However, combining boosting and least angle regression with factorization improves accuracy only sometimes. Additionally, although $r_{PC,IC}$ (Bai & Ng, 2002) more accurately identifies the number of underlying factors than a selection based on AIC and SIC, this only translates to accuracy gains occasionally. Still, both methods are preferred to always include the first three factors. When analyzing dimensionality reduction techniques it is concluded that PCA and Kernel PCA are preferred over SPCA and ICA for the one-step-ahead forecasts constructed in this paper.

Contents

1 Introduction

At the heart of every economic policy decision are forecasts of future economic conditions. Whether it is central banks projecting inflation to set their interest rates, or governments projecting GDP growth to adjust their fiscal policies, forecasts always play an important role. These forecasts are based on a large variety of different economic variables observed over time.

As time passes time-series datasets automatically grow along the time axis T. However, as Donoho et al. (2000) note, with increasing efficiency more and more information is being collected and stored, leading to an explosion of available variables which enter datasets along the predictor axis N. This growth of variables means the dimensionality of datasets increases, which can pose problems for classic statistical techniques (Donoho et al., 2000). Especially, low-frequency economic datasets can become difficult to analyze if T << N.

But hand in hand with the growing datasets, also the literature and methodology to analyze 'Big Data' is growing. Different methodologies for handling high-dimensional datasets have been developed and are constantly being refined. One approach is to use feature selection, for which the goal is to identify a subset of predictors from all the available variables as described in (Mitra et al., 2002) and (Kira & Rendell, 1992). Another approach is to project the dataset into lower dimensionality. A popular technique to reduce dimensionality in such a way is Principal Component Analysis. On the other hand, models such as a Least Angle Regression (Efron et al., 2004) use regularization, which penalizes using many predictors, and therefore selects only the most important predictors automatically. Many other techniques exist and most of them can be employed together. This can make an overview of all available methodologies and their respective performance challenging.

Kim & Swanson (2018) tried to give such an overview, by comparing the forecasting performance of 14 different models for low-frequency macroeconomic data. They analyze different dimensionality reduction techniques, such as PCA, ICA, SPCA, machine learning, variables selection, and shrinkage methods independent from each other and jointly. The authors find that factor models outperform the autoregressive-type benchmark models. Additionally, ML, variable selection, and shrinkage methods are better when combined with factor models. Also, they find that PCA is usually preferred over ICA and SPCA for longer forecasting horizons, due to its robustness, but models based on ICA and SPCA are better when predicting over shorter forecasting horizons. To test whether these findings can be generalized to other data the following two research hypothesis are formulated:

H1: Do factor-based models outperform forecasting accuracy of AR-type models when GDP growth rates are forecasted for 52 African countries?

H2: Does combining machine learning such as boosting and least angle regression with factor models improve performance of these models when forecasting GDP growth rates of 52 African countries?

As indicated by the two hypotheses, the results of Kim & Swanson (2018) will be replicated on a dataset that contains the yearly GDP growth rates of 52 African countries. For the results of Kim & Swanson (2018) the number of factors to include after applying PCA, ICA or SPCA is selected according to the methodology of Bai & Ng (2002). Since factor-based models were shown to be among the best, it is of interest to test how other methodologies for factor selection impact the forecasting performance of these models. Therefore, the following hypothesis is formulated as an extension to the findings of Kim & Swanson (2018):

H3: How does the factor selection methodology impact forecasting accuracy of factor-based models?

On the other hand, Cao et al. (2003) show that the non-linear Kernel Principal Component Analysis (KPCA) in some use cases can deliver better results than the linear PCA or ICA. Since the findings of Kim & Swanson (2018) also show that the preference of PCA, ICA, or SPCA is situation-dependent, including KPCA seems like an interesting extension, which is translated to hypothesis four:

H4: How do KPCA factor-based models compare to PCA, ICA, and SPCA-based models?

By adding to existing literature through the extensions and testing the generalizability of Kim & Swanson (2018), this paper has scientific relevance. At the same time, analyzing the accuracy of GDP growth forecasts is important in a practical way. Accurate economic forecasts are critical for every policy decision that can have an enormous impact on a significant amount of people, highlighting the social relevance of this research.

The rest of this thesis is structured as follows. Section 2 summarizes relevant literature, Section 3 describes the African GDP dataset, and Section 4 covers the Methodology. Section 5 presents the results and which are summarized and discussed in Section 6. Note that, the findings of Kim & Swanson (2018) are not replicated due to the missing dataset. Instead, a simulation study explained in section 4.4 is conducted.

2 Literature Review

2.1 Econometric Literature

Many different methodologies for handling high-dimensional datasets have been developed. Most of the earlier research focused on diffusion index (DI) models as proposed by Stock & Watson

(1998). The idea behind diffusion index models is that a small set of common underlying factors drives the variability in a dataset. Therefore, when constructing diffusion index models, first the underlying factors are estimated from the data, and then those factors instead of the original variables are used in a least-squares regression. As Kim & Swanson (2013) note, the factors are most commonly calculated by PCA. The in general good forecasting accuracy of DI models has been analyzed empirically and theoretically. For example, Stock & Watson (2002) conclude that PCA-based DI models outperform autoregressive benchmark models when predicting industrial production from 149 variables. On the other hand, Bai & Ng (2006) develop a theoretical framework for the use of DI models. They show that the least-squares estimates from factor augmented regressions are consistent and give a methodology for constructing forecast confidence intervals.

When building factor-based models, there is a wide variety of factor estimation and selection techniques. Next to PCA, recent papers such as Kim & Swanson (2018) and Cao et al. (2003) use ICA, SPCA, and KPCA as other ways to estimate the factors. Kim & Swanson (2018) find that for shorter forecasting horizons SPCA and ICA-based factor models deliver better forecasts than PCA-based models and the opposite is true for longer horizons. On the other hand, Cao et al. (2003) find that KPCA performs best, followed by ICA and PCA. Bai & Ng (2002) in their research on determining the correct number of factors show that the common way of selecting factors based on AIC or SIC often overestimates the number of true underlying factors. Instead, they propose new criteria which show better accuracy in their simulation study. In contrast to this, Stock & Watson (2002) find that most forecasting gains come from the first two to three factors and a PCA model which always includes three factors outperforms all other models, including a model where the number of factors is determined according to Bai & Ng (2002).

Next to diffusion index models, other forecasting methodologies such as combination models, LASSO regression, least angle regression, the elastic net, and boosting were shown to produce accurate forecasts in high dimensional datasets. The good performance of combination models which aggregate forecasts from other models is highlighted in Stock & Watson (2004). Stock & Watson (2004) forecast quarterly GDP growth for 7 countries from a set of up to 73 predictors and show that combination forecasts can deliver more accurate forecasts than AR-type benchmark models. Most interestingly, they show that the best and most stable results are achieved for simple weighting schemes such as the arithmetic mean. Bai & Ng (2008) improve the forecasting accuracy of factor-based inflation forecasts by calculating factors from a subset of variables. For the variable selection, LASSO regression, Least Angle regression (LARS), the elastic net, and statistical tests are used. According to Kim & Swanson (2018), one can think

of LASSO, LARS, and the elastic net as penalized regressions with a function of parameters as penalty. When applying these methods, the coefficients of some variables converge to 0 and, therefore, these methods can be also used for variable selection.

Boosting which combines forecasts from simpler models, so-called weak learners, can be employed similarly. Bai & Ng (2009) use boosting for variable selection and improve the accuracy of subsequent 12-month ahead prediction of several macro variables. However, Bai & Ng (2009) note that boosting is sensitive to the data generating process since for example boosting of factors is preferred over boosting of variables only when there is a strong factor structure. Also, Bai & Ng (2008) conclude that the best model is situation-dependent and that the preselected subset of variables changes over time and across predicted variables.

More recent literature such as Kim & Swanson (2018) and Stock & Watson (2012) compare all of these methodologies jointly on even bigger datasets. Stock & Watson (2012) forecast eight US macroeconomic variables from 215 predictor variables for three different forecast horizons with 20 different models. In accordance with Bai & Ng (2008) and Bai & Ng (2009) they find big differences in model performance between the eight variables. Additionally, Stock & Watson (2012) conclude that there are only a few sources of variability as at most six factors account for most of the variation in the dependent variables. Similar to Stock & Watson (2002), the authors also find that including autoregressive terms in the factor-based models offers little or no benefit (Stock & Watson, 2012). Kim & Swanson (2018) compare even more models based on MSFE when predicting eleven US macro variables. Kim & Swanson (2018) show that factor-based models outperform AR-type models, and model averaging techniques such as mean forecasts are best one-third of the time. They also conclude that machine learning and variable selection methods combined with factorization as in Bai & Ng (2008) and Bai & Ng (2009) outperform their pure application. However, also the results of Kim & Swanson (2018) show that different models are preferred between the eleven macro variables.

2.2 Use of Factor Models in Economics

The idea and popularity of diffusion index models can also be seen in classic economic literature. Especially, for asset returns, many theoretical frameworks have been developed that model these returns through a small set of factors. The Capital Asset Pricing Model (CAPM) of Sharpe (1964) was one of the first to do so and models asset returns by a single factor, the market return. According to the CAPM, the expected return of an asset is solely determined by the exposure to market risk. As an alternative to the CAPM, Ross (1976) developed Arbitrage Pricing Theory (APT). In APT, asset returns are driven by multiple undefined factors.

Based on the APT, Fama (1992) developed a three-factor model including the market return, a company size factor, and a value factor based on book-to-market ratios. Hence, they extended the CAPM by two additional factors. Later this was refined to a five-factor model with additional bond-market factors for maturity and default risk as described in Fama & French (1993). The CAPM was also extended to include consumption by Breeden (1979) and became the Consumption CAPM (CCAPM).

3 Data

For the extension, a dataset containing yearly GDP observations of 52 African Countries ranging from 1960 to 2015 (N=52, T=56) is used. From the GDP data for each country i the log GDP growth rates $Y_{i,t}$ are calculated as demonstrated in Equation 1. Since the growth rates cannot be calculated for the first observation, the final sample period is 1961 - 2015 (N=52, T=55). Table A1 in Appendix A shows the descriptive statistics of this dataset.

$$Y_{i,t} = 100 * log(\frac{GDP_{i,t}}{GDP_{i,t-1}}), \quad i = 1, ..., 52, \quad t = 2, ..., 56 \tag{1}$$

From the descriptive statistics, it is concluded that there are large differences in terms of GDP growth between the 52 countries. The log GDP growth of countries such as Somalia seems to be normally distributed with a moderate mean growth of 2%, skewness of -0.463, and kurtosis of 3.033. On the other hand, Rwanda shows a higher average growth of 3.9% but with a significant negative skewness of -4.121 and big swings as seen by the large standard deviation of 12.163 and kurtosis of 25.275. In general, we find a large spread in all statistics. The sample includes stagnating as well as high growth economies as shown by the mean log growth rate ranging from -0.861 (Libya) to 10.168 (Equatorial Guinea). Also, the stability of growth differs as depicted by the standard deviation range of 1.576 (Guinea) to 19.713 (Libya). Overall, this relatively heterogeneous sample could lead to varying model accuracy over the different countries. Therefore, it will be interesting to compare the average forecasting performance to the forecasting performance of individual countries as will be explained in Section 4.7.

4 Methodology

4.1 Dimensionality reduction

The idea behind dimensionality reduction is to meaningfully reduce a high dimensional dataset to a lower dimension (Van Der Maaten et al., 2009). If we denote the African GDP dataset as

a [T x N] = [55 x 52] matrix X[1], then this idea can be explained by Equation 2 taken from Kim & Swanson (2018).

$$X = F\Lambda' + e \tag{2}$$

F represents a [T x r] matrix of factors, Λ is an [N x r] coefficient matrix, and e a [T x N] disturbance matrix. The goal of dimensionality reduction techniques is to find a matrix F with r << N = 52 but that still captures most variation in X. F is not unique and depends on the algorithm that is used to calculate it. In this paper, F is calculated by Principal Component Analysis (PCA), Independent Component Analysis (ICA), Sparse Principal Component Analysis (SPCA), and Kernel Principal Component Analysis (KPCA).

4.1.1 PCA

Principal Component Analysis is one of the oldest and most popular dimensionality reduction techniques (Abdi & Williams, 2010). When applying PCA, orthogonal factors which are called principal components are extracted from the initial dataset. The principal components are linear combinations of the original variables. The first component captures as much variation as possible, while the second tries to do the same but under the condition that it is orthogonal to the first. Similarly, all other principal components are computed (Abdi & Williams, 2010).

$$X = U\Delta V' \tag{3}$$

$$F = U\Delta \tag{4}$$

$$F = U\Delta = U\Delta VV' = XV \tag{5}$$

To see how these factors can be obtained denote the singular value decomposition (SVD) of the dataset X, according to Wall et al. (2003), as in Equation 3. If the dataset X is of size [N x T] with rank L, then the left singular matrix U has size [N x L] and the right singular matrix V has size [L x T]. Δ is then a diagonal [L x L] matrix of singular values.

From Equation 3 then the principal components or factors F can be constructed according to Equation 4. Note that since it is assumed $VV' = I$, where I denotes an identity matrix, F can also be expressed in terms of X and V. Which shows that V represents Λ in the general Equation 2 for PCA.

[1]Note that $Y_i = X_i$ since the columns of X represent the log GDP growth of the 52 different countries over time respectively.

$$X'X = V\Delta^2V' \tag{6}$$

$$U = XV\Delta^{-1} \tag{7}$$

To construct F, one can use the property in Equation 6. Once V and Δ have been obtained, U can be calculated as displayed in Equation 7, where the property that $V^{-1} = V'$ is used.

4.1.2 ICA

Unlike principal components, factors obtained by Independent Component Analysis are assumed to be independent and not ordered by their variances. Similar to the V in PCA Equation 5 the goal is to find a demixing matrix Ψ such that the ICA factors can be calculated as depicted in Equation 8. As Kim & Swanson (2013) note, ICA assumes that the dataset X consists of a statistical independent source data S which is weighted by Ω as seen in Equation 8.

$$F = X\Psi = S\Omega\Psi \tag{8}$$

Note that if we set $\Psi = V$, we would get the same factors as for the PCA. To obtain the demixing matrix Ψ and calculate the ICA factors the "Fast ICA" algorithm proposed by Hyvärinen & Oja (2000) is chosen.

4.1.3 SPCA

One problem with PCA is that, since the factors are linear combinations of all original variables, the factor loadings coefficients are usually non-zero, which makes the interpretation of the different factors difficult. Zou et al. (2006), therefore, developed Sparse Principal Component Analysis which tries to work around this problem by penalizing for non-zero factor loadings. This way, ideally SPCA factors are constructed only from a subset of the most important variables. Similar to Kim & Swanson (2018), this paper follows the methodology of Zou et al. (2006) to obtain the SPCA loadings.

4.1.4 KPCA

Kernel PCA can be seen as a non-linear extension to PCA (Hoffmann, 2007). The idea behind KPCA is to project data that is non-linearly separable into a higher dimension that allows linear separation. The linear principal components that are calculated in the higher dimensional feature space then become non-linear in the original feature space. This method was originally

proposed by Schölkopf et al. (1997) and this paper follows the same proposed methodology to obtain KPCA factors.

4.2 Determining the number of factors r

All of the proposed dimensionality reduction methods return a square matrix containing all calculated factors as columns F_i. Instead of using all factors F_i, the r most important factors will be selected for the construction of the forecasts. There are different methodologies to calculate the best number of factors to include. Let $F^r = [F_1, F_2, ..., F_r]$ denotes the set of selected factors from all available factors. One way to get F^r would be to regress X on F^r as in Equation 2 for $r = 1, 2, ..., rmax$, and then choose the r* that minimizes AIC(r) or SIC(r) of these regressions. However, as Bai & Ng (2002) note, AIC(r) and SIC(r) tend to overestimate the number of factors that should be included. Instead, the authors propose using PC_{p1}, PC_{p2}, IC_{p1} and IC_{p2} as selection criteria instead.

$$PC_{p1}(r) = V(r, \widehat{F}^r) + r\widehat{\sigma}^2(\frac{N+T}{NT})ln(\frac{NT}{N+T}) \tag{9}$$

$$PC_{p2}(r) = V(r, \widehat{F}^r) + r\widehat{\sigma}^2(\frac{N+T}{NT})ln(C_{NT}^2) \tag{10}$$

$$IC_{p1}(r) = ln(V(r, \widehat{F}^r)) + r(\frac{N+T}{NT})ln(\frac{NT}{N+T}) \tag{11}$$

$$IC_{p2}(r) = ln(V(r, \widehat{F}^r)) + r(\frac{N+T}{NT})ln(C_{NT}^2) \tag{12}$$

These criteria are calculated according to Equation 9 to 12. $V(r, \widehat{F}^r)$ are the squared residuals of regressing X on $\widehat{F}^r$ divided by T and $C_{NT}^2 = min(T, N)$. As Bai & Ng (2002) note, all of the proposed criteria converge in infinite samples but have different properties for finite samples.

$$r_{AIC,SIC} \approx \frac{arg_r minAIC(r) + arg_r minSIC(r)}{2} \tag{13}$$

$$r_{PC,IC} \approx \frac{arg_r minPC_{p1} + arg_r minPC_{p2} + arg_r minIC_{p1} + arg_r minIC_{p2}}{4} \tag{14}$$

Since r, the number of factors that are included can impact forecasting accuracy, this paper will follow three separate methodologies for determining r. First, r is selected according to AIC(r) and SIC(r) by rounding the average suggested number of factors from both measures as depicted in Equation 13. Additionally r is calculated based on PC_{p1}, PC_{p2}, IC_{p1} and IC_{p2} as proposed by Bai & Ng (2002). The average of all four suggested r is taken and rounded to the nearest integer as seen in Equation 14. For the calculations of the criteria we set $rmax = 15$. Finally, both methodologies are also compared to fixing $r_3 = 3$, which showed the best results for Stock & Watson (2002).

4.3 Forecasting methodology

Following Kim & Swanson (2018), the h-step ahead growth rate forecast of country i $Y_{i,t+h}$ are constructed as in Equation 15. $W_{i,t}$ is a vector containing additional variables, such as the lagged dependent variable. F_t is a [1 x r] vector of factors which are calculated and selected as explained in Section 4.1 and 4.2. $\beta_{i,W}$ and $\beta_{i,F}$ are the coefficient vectors calculated by Ordinary Least Squares (OLS). $\epsilon_{i,t+h}$ is the disturbance term.

$$Y_{i,t+h} = W_{i,t}\beta_{i,W} + F_{i,t}\beta_{i,F} + \epsilon_{i,t+h}, \quad i = 1, ..., 52, \quad t = 1, ...55. \tag{15}$$

This paper will focus on one-step-ahead forecasts, such that h = 1. Forecasts are constructed on an expanding- and rolling window basis. For both methods, the initial estimation sample is the first $T_{est.} = 40$ years (1961 - 2000). After constructing the one-step-ahead forecast for 2001 the expanding window sample increases to $T_{est.} = 41$ observations, while the rolling window estimation period rolls forward by one year keeping the 40 most recent observations. In total there are therefore 2 (expanding window, rolling window) x 4 (PCA, ICA, SPCA, KPCA) x 3 (r_3, $r_{AIC,SIC}$, $r_{PC,IC}$) = 24 different forecast settings.

Similar to Kim & Swanson (2018), all estimation calculations are redone for each new forecast. This includes the calculation and selection of factors $F_{i.t}$, re-estimation of all regression coefficients such as $\beta_{i,W}$ and $\beta_{i,F}$, all lag order selections such as of the dependent variable included in $W_{i,t}$, and all other machine learning, variable selection and shrinkage methods explained below.

4.4 Simulation Study

To test how the factor selection methodologies and model performance depend on the number of variables N and observations T, six datasets are simulated. For this first R factors F_i are constructed from an autoregressive process with lag p = 1, as shown in the following Equation:

$$F_{i,t} = \phi_i F_{i,t-1} + \epsilon_{i,t}, \quad i = 1, ..., R, \quad t = 1, ..., T \tag{16}$$

$F_{i,0} = 0$ for all i, $\epsilon_{i,t} \sim IID$ N(0,1), and ϕ_i, T, and R vary over the different data generating processes. Next the [T x N] dataset X and the [T x 1] dependent variable Y are generated from the factors F_i, according to Equation 17 and 18 respectively. The factor loadings $\lambda_{j,i} \sim IID$ N(0,1), likewise $\epsilon_{j,t} \sim IID$ N(0,1), and N varies over the different data generating processes.

$$X_j, t = \lambda_{j,1}F_{1,t} + ... + \lambda_{j,R}F_{R,t} + \epsilon_{j,t}, \quad j = 1, ..., N, \quad t = 1, ..., T \tag{17}$$

The factor loadings for Y_t are set equal to one and the disturbance term $\epsilon_t \sim IID$ N(0,1).

$$Y_t = F_{1,t} + ... + F_{R,t} + \epsilon_t, \quad t = 1, ..., T \tag{18}$$

The settings for all six different data generating processes can be found in Table 1. Two different sample sizes of [50 x 50] and [100 x 100] are chosen, to test the accuracy of $r_{AIC,SIC}$ and $r_{PC,IC}$ in a sample the size of the African GDP dataset and in a slightly bigger one. The number of factors is varied to include either one, four, or eight, to test if this affects the factor selection methodologies. Since the focus of this paper is on forecasting, one step ahead forecast $\widehat{Y}_{t+1}$ are constructed from X_t.

Table 1: Specification for six different data generating processes

	T	N	R	ϕ_i
DGP1	50	50	1	[0.9]
DGP2	50	50	4	[0.9, 0.8, 0.7, 0.6]
DGP3	50	50	8	[0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2]
DGP4	100	100	1	[0.9]
DGP5	100	100	4	[0.9, 0.8, 0.7, 0.6]
DGP6	100	100	8	[0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2]

4.5 Forecasting models depending on PCA, ICA, SPCA and KPCA

The models discussed in this section are of the form presented in Equation 15 and for all models $F_{i,t}$ is calculated by PCA, ICA, SPCA, and KPCA. Therefore, models will differ depending on the factor estimation technique. The models that will be compared include: Component Regression, Factor augmented regression, Boosting, and Least Angle Regression.

$$\widehat{Y}^{CA}_{i,t+1} = \widehat{\alpha} + \widehat{\gamma} F_{i,t} \tag{19}$$

The Component Regression (**CA**) in Equation 19 is a deviation from Kim & Swanson (2018) who include a principal component regression and do not calculate this model for ICA, SPCA and KPCA.

$$\widehat{Y}^{FAAR}_{i,t+1} = \widehat{\alpha} + \sum_{j=1}^{p} \widehat{\phi}_j Y_{i,t+1-j} + \widehat{\gamma} F_{i,t} \tag{20}$$

The Factor augmented autoregression (**FAAR**) model presented in Equation 20 is a combination of an AR(p) and CA model. Like with those models, first, the lag order p is selected

using SIC, and then the regression on $F_{i,t}$ takes place.

$$\widehat{Y}_{i,t+1}^{Boosting(F)} = F_m(X_{i,t}) = \sum_{m=1}^{M} h_m(X_{i,t}) \tag{21}$$

$$F_m(X_i) = F_{m-1}(X_i) + h_m(X_i) \tag{22}$$

$$h_m = arg_h min \sum_{i=1}^{n} l(Y_i, F_{m-1}(X_i) + h(X_i)) \tag{23}$$

Boosting (**Boosting(F)**) is represented by a Gradient Boosting Regression Tree model (GBRT). Hastie et al. (2009) explain the idea behind boosting as combining forecasts from multiple weak learners h_m as seen in Equation 21. The weak learners are simpler models and are decision tree regressions for GBRT. The hyperparameter M denotes the number of boosting stages. As can be seen from Equation 22, the GBRT algorithm is constructed iteratively. h_m is calculated according to Equation 23, where l() denotes a loss function that is specified as a hyperparameter. Similar to 'Specification 1' of Kim & Swanson (2013), this model is estimated on $F_{i,t}$ after the factors have been calculated by either PCA, ICA, SPCA, or KPCA. This specification is chosen as the results of Kim & Swanson (2013) show that 'Specification 1' was among the best specifications for expanding window h=1 forecasts. The hyperparameters of this model are: loss function = least squares, learning rate = 0.1, boosting stages = 100.

A Least Angle Regression (**LARS(F)**) is performed according to the methodology proposed by Efron et al. (2004). As with the Boosting, the Least Angle Regression is performed on $F_{i,t}$. Additionally, 5-fold cross-validation is used to avoid overfitting the model.

Kim & Swanson (2018) compute additional models which are not replicated in this paper. The models that are excluded were the worst-performing models for the h=1 forecast horizon. More specifically, the following models given together with their percentage of wins (being the best model) are excluded: bagging (0.00%), bayesian model averaging with two different g-prior (1.95%, 1.62%), ridge regression (0.65%), elastic net (2.60%), and non-negative garotte (0.32%).

4.6 Baseline models

Next to the models already discussed, also the following benchmark models are included. These models are considered benchmarks since, except for the mean model, they do not use $F_{i,t}$ and, therefore, give the same forecast for PCA, ICA, SPCA, and KPCA.

$$\widehat{Y}_{i,t+1}^{AR} = \widehat{\alpha} + \sum_{j=1}^{p} \widehat{\phi}_j Y_{i,t+1-j} \tag{24}$$

AR(p) (**AR**) model as depicted in Equation 24 where the number of lags p is selected according to the SIC.

$$\widehat{Y}_{i,t+1}^{ARX} = \widehat{\alpha} + \sum_{j=1}^{p} \widehat{\phi}_j Y_{i,t+1-j} + \widehat{\beta} Z_{i,t} \tag{25}$$

ARX(p) (**ARX**) model as depicted in Equation 25 where first the number of lags of the dependent variable p is determined by SIC. Then, similar to Kim & Swanson (2018) the first lag of each variable in the dataset X is added to the regression iteratively. If the adjusted R^2 of the regression improves by at least 0.01 the variable is kept and added to $Z_{i,t}$. After this has been done for all variables, the same process is repeated for the second lag and repeated until the sixth lag of each variable.

$$\begin{aligned} \widehat{Y}_{i,t+1}^{ADL,k} &= \widehat{\alpha} + \sum_{j=1}^{p_{k,y}} \widehat{\phi}_j Y_{i,t+1-j} + \sum_{j=1}^{p_{k,x}} \widehat{\beta}_j X_{k,t+1-j}, \quad for \quad k = 1, ..., N \\ \widehat{Y}_{i,t+1}^{CADL} &= \frac{1}{N} \sum_{k=1}^{N} \widehat{Y}_{i,t+1}^{ADL,k} \end{aligned} \tag{26}$$

Combined bivariate autoregressive distributed lag model (**CADL**) as proposed by Stock & Watson (2012) and shown in Equation 26. Similar to Kim & Swanson (2018) 52 individual ADL models are constructed for each of the 52 X_k variables for this model. For these models, first, $p_{k,x}$, the lag order of X_k is determined by SIC and then $p_{k,y}$ the best lag order of the dependent variables is calculated also by SIC. The average of these 52 individual ADL models is then the combined bivariate ADL model.

Boosting (**Boosting(X)**), as discussed in Section 4.5, is also applied to the full dataset X making the forecast independent of the choice of dimensionality reduction technique. This model can be compared to the boosting model under 'Specification 3' of Kim & Swanson (2018).

Similarly, a Least Angle regression (**LARS(X)**) is also performed on the full dataset X.

The final model is the arithmetic mean (**Mean**) of all other nine forecasts. Therefore, unlike the other benchmark models, the forecasts from this model will differ for PCA, ICA, SPCA, and KPCA. Table A2 in Appendix A gives an overview of all 10 models.

4.7 Model comparison

To compare the forecasting accuracy of the different models, forecasts for the GDP growth rates of all 52 countries are constructed. Then the Mean Squared Forecast Error ($MSFE_i$) is calculated for the forecasts of each country i according to Equation 27. $\widehat{e}_{i,t}$ denotes the forecast

error for country i computed as $Y_{i,t} - \widehat{Y}_{i,t}$. The average MSFE over all countries is computed and is scaled by dividing by the MSFE of the AR(p) benchmark.

$$MSFE_i = \frac{1}{\#(T_{forecast})} \sum_{t \in T_{forecast}} \widehat{e}_{i,t}^2 \tag{27}$$

As noted in Section 3 the dataset seems to be heterogeneous, and growth characteristics differ for countries. Therefore, also the 'wins' per country will be counted. Win refers to a model having the lowest $MSFE_i$ for a given country i.

5 Results

5.1 Simulated Dataset

First, the results of the simulated dataset are presented. Table 2 gives the mean and accuracy of $r_{PC,IC}$ and $r_{AIC,SIC}$, as well as the best model (lowest MSFE) for all six data generating processes and different dimensionality reduction techniques. This table shows that the $r_{PC,IC}$ does well, attaining 100 percent accuracy most of the time. There are only two exceptions. First of all, when applying ICA, the number of selected factors is too high. This can be explained by convergence problems of the proposed 'Fast ICA' algorithm. However, this impacts the forecasting performance of factor-based ICA models only partially as will be shown. Next to the ICA, $r_{PC,IC}$ struggles to identify the 8 underlying factors of DGP3. When applying PCA or KPCA on average 7.667 factors are selected with an accuracy of 0.667, while for SPCA the correct 8 factors were identified all of the time. However, when the sample size is increased to [100 x 100] as in DGP6, $r_{PC,IC}$ has an accuracy of 100 percent, except for the mentioned ICA issue. Nevertheless, the sample size of DGP3 resembles the African GDP dataset of size [52 x 55]. From these results, it is concluded, $r_{PC,IC}$ is accurate and might only have troubles when many underlying factors influence the GDP growth of those countries.

On the other hand, as suggested by Bai & Ng (2002) $r_{AIC,SIC}$ overestimates the number of underlying factors. Except for KPCA and when there is only 1 true factor (DGP1 and DGP4), the mean factor selected is above the true value. For KPCA $r_{AIC,SIC}$ always selects the correct number of factors. Even for DGP3 where $r_{PC,IC}$ failed. But also for PCA and SPCA are factors selected correctly sometimes as can be seen from the accuracy range of 0.222 to 0.889 and the mean values are close to the true R.

Overall, Least Angle Regression and Boosting seem to be the best models. Only for DGP6, which contains many observations and 8 factors, the Mean model makes the best forecasts for both factor selection methodologies. In general, the same model is best for both $r_{PC,IC}$ or

Table 2: Best model and accuracy of $r_{PC,IC}$ and $r_{AIC,SIC}$ for one step ahead expanding window forecasts of the 10 last observation in six simulated datasets for PCA, ICA, SPCA, and KPCA

			$r_{PC,IC}$			$r_{AIC,SIC}$		
	R	Dim. Red.	Mean	Accuracy	Best Model	Mean	Accuracy	Best Model
DGP1	1	PCA	1.000	1.000	**LARS(F)**	1.000	1.000	**LARS(F)**
	1	ICA	11.222	0.111	LARS(X)	11.889	0.000	LARS(X)
	1	SPCA	1.000	1.000	**LARS(F)**	1.000	1.000	**LARS(F)**
	1	KPCA	1.000	1.000	**LARS(F)**	1.000	1.000	**LARS(F)**
DGP2	4	PCA	4.000	1.000	Boosting(F)	4.444	0.778	**Boosting(F)***
	4	ICA	13.333	0.000	CA*	18.778	0.000	LARS(X)
	4	SPCA	4.000	1.000	Boosting(F)*	9.778	0.222	CA
	4	KPCA	4.000	1.000	Boosting(F)*	4.000	1.000	Boosting(F)
DGP3	8	PCA	7.667	0.667	Boosting(F)	8.111	0.889	Boosting(F)*
	8	ICA	11.111	0.111	Boosting(X)	18.111	0.000	Boosting(X)
	8	SPCA	8.000	1.000	Boosting(X)	10.000	0.333	Boosting(X)
	8	KPCA	7.667	0.667	LARS(F)	8.000	1.000	**Boosting(F)***
DGP4	1	PCA	1.000	1.000	**LARS(X)**	1.000	1.000	**LARS(X)**
	1	ICA	11.222	0.000	**LARS(X)**	15.667	0.000	**LARS(X)**
	1	SPCA	1.000	1.000	**LARS(X)**	1.000	1.000	**LARS(X)**
	1	KPCA	1.000	1.000	**LARS(X)**	1.000	1.000	**LARS(X)**
DGP5	4	PCA	4.000	1.000	LARS(X)	4.444	0.667	LARS(X)
	4	ICA	16.444	0.000	LARS(F)	19.000	0.000	**CA***
	4	SPCA	4.000	1.000	LARS(X)	4.556	0.889	LARS(X)
	4	KPCA	4.000	1.000	LARS(X)	4.000	1.000	LARS(X)
DGP6	8	PCA	8.000	1.000	Mean	8.333	0.778	**Mean***
	8	ICA	15.778	0.000	CADL	19.333	0.000	CADL
	8	SPCA	8.000	1.000	Mean	11.000	0.333	Mean*
	8	KPCA	8.000	1.000	Mean	8.000	1.000	Mean*

Note: * marks the better model per row, no * means both model were equally accurate. The best model for each DGP is given in bold.

$r_{AIC,SIC}$. The only four exceptions are DGP2 ICA, DGP2 SPCA, DGP3 KPCA, DGP5 ICA, which all have a bigger difference in the average number of selected factors.

Furthermore, it is observed that for the smaller datasets (DGP1-3) 17/24 best models are factor-based. For the larger datasets (DGP4-6), this number drops to 2/24 (8/24 if mean is also included). For the larger dataset LARS(X) delivers the best forecasts most of the time and Boosting never does. This indicates that the advantage of combining dimensionality reduction with Least Angle Regression and Boosting disappears in larger datasets. It will be interesting to see what conclusion can be drawn for the African GDP dataset that is of 'small' size.

Usually, the same model delivers the best forecasts for PCA, SPCA, and KPCA, but a different model for ICA. For ICA, in DGP1 and DGP3 sample-based models are selected (LARS(X) instead of LARS(F) and Boosting(X) instead of Boosting(F) respectively), which is probably due to the ICA convergence issue resulting in bad factors. This could also explain why for DGP6 CADL model is better than the mean. On the other hand, for DGP2 and DGP5 which both contain 4 true factors, factor-based models deliver the best ICA forecasts (CA and LARS(F) for $r_{PC,IC}$, and LARS(X) and CA for $r_{AIC,SIC}$ respectively). In the case of the DGP5, the CA model of $r_{AIC,SIC}$ even is the best overall model for that data generating process.

If the best models from $r_{PC,IC}$ and $r_{AIC,SIC}$ are compared directly, 3/24 times $r_{PC,IC}$ models outperform their $r_{AIC,SIC}$ counterparts. 7/24 times the opposite holds true but for a majority of 14/24 times the best model of both methodologies has equal precision. The overall best model for each of the six data generating processes comes from $r_{AIC,SIC}$ 4/6 times. For DGP1, LARS(F) delivered the same accuracy for PCA, SPCA, and KPCA, across both methodologies. For DGP4 the LARS(X) model delivered the most accurate forecasts and was, therefore, the best model, independent of the dimensionality reduction technique. As a result, it is concluded that the better accuracy of $r_{PC,IC}$ does not seem to significantly benefit forecasting accuracy.

Finally, the best models are not dominated by a specific dimensionality reduction technique. For DGP1, LARS(F) delivered the same forecasts for PCA, SPCA, and KPCA, which would mean that all three methodologies calculated the same or a very similar factor. PCA-based models delivered the best results for DGP2 and DGP6, KPCA for DGP3, and ICA for DGP5. For DGP4 no dimensionality reduction managed to produce a model that outperformed LARS(X).

Table A3 in Appendix A shows the same results as in Table 2 for rolling window forecasts. Mostly the same results hold also for the rolling window forecasts. $r_{PC,IC}$ identifies the number of factors more accurately than $r_{AIC,SIC}$. The same ICA issue is observed and usually, another model is picked then for PCA, SPCA, and KPCA. Additionally, factor models are also preferred for the smaller datasets but for DGP4-DGP6 more benchmark models manage to give the best

forecasts. On the other hand, LARS and Boosting do not dominate as much as for the expanding window, with FAAR and CA appearing more often. Also $r_{PC,IC}$ models do slightly better than $r_{AIC,SIC}$ in terms of direct comparison and being the overall best models for a DGP.

In summary, the results of the simulated dataset indicate that $r_{PC,IC}$ does a better job at identifying the true number of underlying factors than $r_{AIC,SIC}$ as suggested by Bai & Ng (2002). However, this did not translate to better forecasting accuracy and usually, the same model is the best for both factor selection methodologies. Least Angle Regression and Boosting give the best forecasts most of the time. For the smaller DGP1-3, combining these models with dimensionality reduction delivers better results but for the larger DGP4-6 LARS(X) dominates all other models. Except for the ICA, which has convergence issues, the dimensionality reduction techniques only seem to have a marginal impact on the best model's forecasting accuracy. The majority of these observations also hold for rolling window forecasts. To test whether these findings can be generalized, a similar analysis is conducted on the African GDP Dataset in the next section.

5.2 Africa GDP Dataset

For the results in this section, forecasts for the log GDP growth rates of all 52 African countries are made and then evaluated by $MSFE$. As described in the data section, the growth statistics differ a lot across the different countries. Therefore, the performance across all countries is evaluated first before giving a more detailed view of the forecasting performance for individual countries and how the factor selection and dimensionality reduction methodologies influence the forecast models.

5.2.1 Full sample results

Table 3 shows the $MSFE$ of the best model relative to the $MSFE$ of the AR(p) benchmark for the different factor selection and dimensionality reduction techniques. The $MSFE$ is calculated from the predictions of all countries. The values of the best model are 0.004 across the whole table because the LARS(X) model delivered the most accurate forecasts for the full sample. The expanding window gain over the AR benchmark is bigger than the rolling window gain, but they are rounded to the same integer. As a result the overall best model always comes from the expanding window, similar to Kim & Swanson (2018) who also find the expanding window forecasts are preferred over rolling window forecasts for the h=1 forecast horizon. The results indicate that the Least Angle Regression without dimensionality reduction does much better than the AR(p) benchmark. The LARS(X) model also was one of the best models for the

simulated datasets. To find a possible explanation for the strong performance of the LARS(X) model, the results are analyzed in more detail.

Table 3: Relative MSFE of best model based on all forecasts under 24 specifications

	Expanding window				Rolling window		
	r_3	$r_{AIC,SIC}$	$r_{PC,IC}$		r_3	$r_{AIC,SIC}$	$r_{PC,IC}$
PCA	0.004	0.004	0.004	PCA	0.004	0.004	0.004
ICA	0.004	0.004	0.004	ICA	0.004	0.004	0.004
SPCA	0.004	0.004	0.004	SPCA	0.004	0.004	0.004
KPCA	0.004	0.004	0.004	KPCA	0.004	0.004	0.004

First of all, to give a better overview of the performance of the other models, Figure 1 and Figure 2 show the ranking (in terms of $MSFE$) of all models for all specifications of the expanding and rolling window forecasts respectively. Both figures show relatively stable rankings as most models remain within ± 1 rank of their average rank over time, except for when ICA is applied. If factors are calculated according to ICA, the factor-based models LARS(F), CA, and FAAR struggle to produce good forecasts. Only Boosting(F) seems to be resilient. Together with all non-factor-based models, the rank of Boosting(F) increases due to the performance dip of the three other factor-based models.

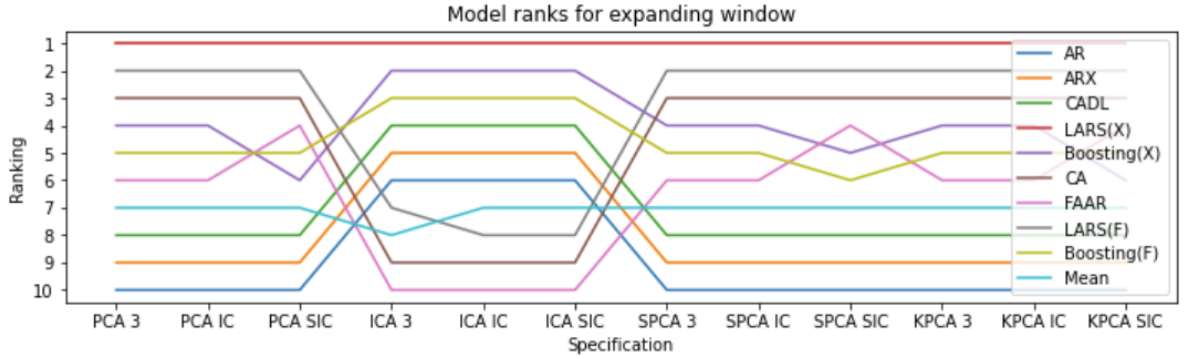

Figure 1: MSFE ranking of models for expanding window forecasts and all specifications

As was clear from Table 3, LARS(X) takes first place for both windows across all specifications. The second-best model is LARS(F) unless factors are calculated by ICA. For the expanding window, the CA forecasts are usually a bit more accurate than the Boosting(X). However, unlike Boosting(X), CA is dependent on the dimensionality reduction technique and does poorly when ICA is applied. For the rolling window CA forecasts are better than Boosting(X) forecasts when the factors are determined by $r_{PC,IC}$ and also when the first three factors of PCA and KPCA are chosen, but worse else. Boosting(F) and FAAR are in the middle field, followed by

the mean forecast. The bottom three ranks go to CADL, ARX, and the AR benchmark. CADL always outperforms ARX, which in turn is always more accurate than the AR model. Since the forecasts from all three models do not differ across the specifications it is clear that the better ranks of these models for the ICA specifications, come from the underperformance of the factor-based models.

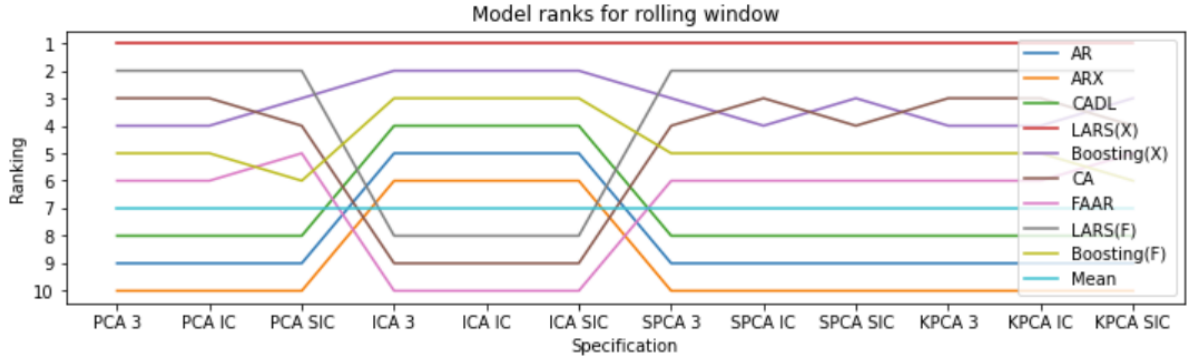

Figure 2: MSFE ranking of models for rolling window forecasts and all specifications

Overall, the full sample results can be summarized in the light of the first two research hypotheses as follows. First of all, factor-based models indeed seem to outperform AR-type models. Except for when ICA is applied, factor-based models such as CA, FAAR, LARS(F), and Boosting(F) tend to be always better than the AR and ARX models. This is in line with findings of Kim & Swanson (2018), Stock & Watson (2002), Stock & Watson (2004), and Stock & Watson (2012).

Secondly, combining Boosting and LARS with dimensionality reduction in the form of PCA, ICA, SPCA, and KPCA does not seem to bring any advantage. The LARS(X) model always outperforms the LARS(F) and the Boosting(X) model also generates more accurate forecasts than Boosting(F) for most of the specifications. This is in contrast to the results of Kim & Swanson (2018), who find that machine learning, variable selection, or shrinkage model are only good when combined with factors. This difference can potentially be explained by the different datasets. The dataset of Kim & Swanson (2018) contains almost three times as many variables and more than ten times more observations than the African GDP dataset. This means that in the African dataset much less information is available to determine any underlying factors. Additionally, the research of Bai & Ng (2009) highlights that boosting the variables is preferred if there is no strong factor structure. The large difference in the African countries might indicate that this is the case.

Finally, the findings can also be related to the results of the simulated dataset. Similar to those results, Boosting and Least Angle Regressions are the best models. As for the larger

datasets, LARS(X) is the overall best model for most cases. Furthermore, the dimensionality reduction and factor selection methodologies only marginally impact the ranking of the different models, except for ICA. Most of the main findings, therefore, hold when looking at average performance. However, since this assessment is only based on the average forecasting accuracy of all 52 countries, further analysis is needed.

5.2.2 Country results

Due to the large difference in countries, the number of wins is analyzed next. This analysis is conducted since the average results could potentially be skewed. For example, a model could potentially produce accurate forecasts for a large number of countries, but be completely off for a few other countries, which could increase the average $MSFE$ disproportionally. Such a scenario could explain why the factor-based and AR-type models, for which the set of explanatory variables changes through renewed dimensionality reduction and variable selection based on SIC, seem to do worse than LARS(X) which has the same set of 52 explanatory variables for each prediction.

Table 4: Model that most often was the lowest MSFE model counted over all 52 countries for PCA, ICA, SPCA, KPCA and different ways to select the number of factors r

	Expanding window				Rolling window		
	r_3	$r_{AIC,SIC}$	$r_{PC,IC}$		r_3	$r_{AIC,SIC}$	$r_{PC,IC}$
PCA	LARS(X)	LARS(X)	LARS(F)[1]	PCA	LARS(X)	LARS(F)[2]	LARS(F)
ICA	LARS(X)	LARS(X)	LARS(X)	ICA	LARS(X)	LARS(X)	LARS(X)
SPCA	LARS(X)	LARS(X)	LARS(X)	SPCA	LARS(X)	LARS(X)	LARS(X)
KPCA	LARS(X)	LARS(X)	LARS(F)[1]	KPCA	LARS(X)	LARS(X)	LARS(F)

Note: [1]together with Boosting(X) and LARS(X), [2]together with LARS(X).

To begin with, Table 4 shows the model which won the most countries for each specification. For example, the LARS(X) of PCA and r_3 indicates that no model won more of the 52 countries than the LARS(X) for that specification. As with the average results, this table is dominated by LARS(X). However, when applying $r_{PC,IC}$ other models win more countries than the LARS(X). For the expanding window forecasts, LARS(F) together with LARS(X) and Boosting(X), are the best models for 13 countries each, when applying PCA or KPCA and using $r_{PC,IC}$ to select the factors. For the rolling window estimation, PCA or KPCA combined with $r_{PC,IC}$ result in LARS(F) being the best model for 16 countries, compared to 12 for the LARS(X). Lastly, also for the rolling window PCA $r_{AIC,SIC}$ specification, LARS(F) is tied with LARS(X) for most

wins with 13 wins each.

In general, it can be observed that whenever the number of factors is determined by $r_{PC,IC}$, LARS(F) wins more countries at the expense of LARS(X). The results of Table 4 suggest that there are some countries for which LARS(X) is not the ideal model. To illustrate this, Table 5 shows the model that was most often the best model for each country with the percentage of wins over the 24 specifications. The results of Table 5 clearly show that the different growth patterns impact which model gives the best forecasts.

Confirming earlier results, LARS(X), LARS(F), and Boosting(X) appearing the most often in Table 5. However, also models that did worse on average beat these three models for some of the countries. For example, the ARX model gives the best growth forecast for the Central African Republic for 12/24 specifications. For Mauritius, the same holds for the CADL model, and for Ghana, the CADL model always gives the most accurate forecasts. Togo's GDP growth is best forecasted by the CA model for half of the specifications.

Not only do the best models differ over the countries, but also the extent of their dominance. For countries such as Burkinafaso, Cabo Verde, Djibouti, Rwanda (Boosting(X)), Equatorial Guinea, Liberia, Seychelles (LARS(X)), and Ghana (CADL) one model always gives the best forecasts. For other countries, the best model choice is dependent on the specifications and, therefore, changes a lot. An example of this is Benin where no model manages to be best for more than a third of the specifications. On the other hand, countries like Egypt, Libya, and Nigeria are best predicted by two models that split the top spot. For Nigeria, depending on the specification, either LARS(F) or LARS(X) gives the best forecasts and the overall wins are split evenly in half.

In summary, as expected, the more in-depth country by country results show that the LARS(X) does not produce the best forecasts for all of the countries and specifications. Table 4 suggests that when using $r_{PC,IC}$ together with PCA or KPCA, then the LARS(F) gives the best forecasts for more countries than the LARS(X). Furthermore, the difference in the GDP growth statistics resulted in a big variety of models present in Table 5. While for some countries the same model gives the best forecast independent of specifications, for most countries the best model changes over the different specifications.

Similarly, Kim & Swanson (2018) find that the best model differs across the eleven macroeconomic variables that are being predicted. For most variables, the best model also changes depending on the specification and forecast horizons. The macroeconomic variables of Kim & Swanson (2018) also consist of many different types of variables. So much like the dataset of this paper, the difference in those variables can explain the variation in best models.

Table 5: Best model per country with percentage of specifications won out of 24

Country	Best Model	% Wins	Country	Best Model	% Wins
algeria	LARS(F)	0.750	liberia	LARS(X)	1.000
angola	LARS(X)	0.833	libya	LARS(X)/Boosting(F)	0.417
benin	LARS(F)	0.333	madagascar	LARS(X)	0.917
botswana	Boosting(X)	0.625	malawi	LARS(X)	0.667
burkinafaso	Boosting(X)	1.000	mali	LARS(F)	0.542
burundi	LARS(F)	0.625	mauritania	LARS(F)	0.500
caboverde	Boosting(X)	1.000	mauritius	CADL	0.500
cameroon	LARS(F)	0.750	morocco	LARS(X)	0.750
car	ARX	0.500	mozambique	Boosting(X)	0.917
chad	LARS(X)	0.500	namibia	LARS(X)	0.500
comoros	LARS(F)	0.542	niger	LARS(F)	0.458
congodr	Boosting(X)	0.667	nigeria	LARS(F)/LARS(X)	0.500
congorepub	LARS(X)	0.542	rwanda	Boosting(X)	1.000
djibouti	Boosting(X)	1.000	saotome	LARS(F)	0.292
egypt	Boosting(X)/LARS(X)	0.500	senegal	LARS(F)	0.500
eqguinea	LARS(X)	1.000	seychelles	LARS(X)	1.000
eritrea	Boosting(X)	0.500	sierraleone	LARS(F)	0.458
ethiopia	Boosting(X)	0.958	somalia	LARS(X)	0.667
gabon	LARS(X)	0.500	southafrica	LARS(X)	0.625
gambia	LARS(F)	0.542	sudan	LARS(X)	0.792
ghana	CADL	1.000	tanzania	Boosting(X)	0.542
guinea	LARS(X)	0.542	togo	CA	0.500
guineabissau	LARS(X)	0.750	tunisia	LARS(X)	0.583
ivorycoast	Boosting(X)	1.000	uganda	Boosting(X)	0.917
kenya	LARS(X)	0.458	zambia	LARS(X)	0.417
lesotho	Boosting(X)	0.375	zimbabwe	LARS(X)	0.750

Also in terms of wins, LARS and Boosting are the best performing models, like in the simulated dataset. The better performance of LARS(F) is also more in line with the results of DGP1-3, which closely resemble the African dataset in size. The more in-depth country results allow refining the answers to the first two research questions. The first hypothesis, that factor-based models outperform AR-type models is further strengthened due to more wins of factor-based models. On the other hand, the results of this section show that under certain conditions, LARS(F) outperforms LARS(X) indicating that there are benefits from a combination with dimensionality reduction. To determine these conditions and to be able to answer research hypotheses three and four, the impact of factor selection methodology and dimensionality reduction technique is analyzed further in the next section.

5.2.3 Factor selection and dimensionality reduction

The results from Table 4 indicated that LARS(F) wins more countries at the expense of the LARS(X) whenever $r_{PC,IC}$ is used. To further examine this, Table 6 shows how often each models delivered the best forecasts for the different factor selection methodologies r_3, $r_{AIC,SIC}$, and $r_{PC,IC}$. The columns add up to 4 * 52 = 208 as the wins are counted for all dimensionality reduction techniques.

Table 6: Wins per model for different ways of selecting the number of factors r counted over all four dimensionality reduction techniques

	Expanding window				Rolling window				Both
Model	r_3	$r_{AIC,SIC}$	$r_{PC,IC}$	Total	r_3	$r_{AIC,SIC}$	$r_{PC,IC}$	Total	Total
AR(p)	0	0	0	0	0	0	0	0	0
ARX(p)	4	4	4	12	0	0	0	0	12
CADL	8	6	6	20	15	13	13	41	61
CA	7	8	23	38	5	11	9	25	63
FAAR	10	3	5	18	10	9	12	31	49
Boosting(X)	51	51	52	154	44	43	44	131	285
Boosting(F)	5	15	7	27	6	15	9	30	57
LARS(X)	86	88	73	247	84	74	69	227	474
LARS(F)	37	33	37	107	38	36	46	120	227
Mean	0	0	1	1	6	7	6	19	20

First of all, the last column of Table 6 gives the total number of wins for each model. This allows to get a complete ranking of all models also in terms of their wins rather than based on

$MSFE$ as done previously. The LARS(X) (474) delivered the most accurate forecasts in most cases, followed by Boosting(X) (285) and LARS(F) (227). The fourth best model in terms of wins is the CA (63), tightly followed by CADL (61), Boosting(F) (57), and FAAR (49). The three bottom places go to Mean forecast (20), ARX (12), and the AR (0) model which never delivered the most accurate forecasts. This ranking, in essence, resembles the ranking based on $MSFE$. Only the CADL model and Boosting(X) do slightly better than in the $MSFE$ ranking. The CADL, FAAR, LARS(F), and Mean models all get the majority of their wins under the rolling window forecast. The ARX, CA, Boosting(X), and LARS(X) do relatively better for the expanding window forecasts.

Next, the wins can be broken down over the different factor selection methodologies, to show their impact on forecasting accuracy. For the expanding window, the biggest difference in the number of wins is seen in the CA, Boosting(F), and LARS(X) models. When using $r_{AIC,SIC}$ instead of r_3, Boosting(F) wins 15 countries instead of 5. Since the wins of most other models remain stable, the Boosting(F) gains are most likely at the costs of FAAR and LARS(F) for which the wins drop by 7 and 4 respectively. Bigger changes are seen when using $r_{PC,IC}$ for the factor selection. Under $r_{PC,IC}$ the number of wins for the CA triple to 23, and the LARS(X) wins drop to 73, well below the wins for the two other methodologies. The FAAR and LARS(F) win increase slightly when compared to the $r_{AIC,SIC}$ wins but remain below or equal to the r_3 wins. The Boosting(F) wins are 7 which is below the 15 of $r_{AIC,SIC}$ but still above the 5 from r_3.

These movements can be summarized in the following patterns. First of all, when using $r_{AIC,SIC}$, which according to Bai & Ng (2002) and the simulation results overestimates the number of true underlying factors, Boosting(F) gains more wins at the expense of FAAR and LARS(F). However, when the supposedly more accurate $r_{PC,IC}$ is used, factor-based models such as CA, FAAR, and LARS(F) gain at the cost of LARS(X) and Boosting(F). This first finding indicates that Boosting(F) is the most robust of the factor-based models, doing better even if factors are selected based on $r_{AIC,SIC}$. This is in line with the results of Figure 1 and 2 which already showed the robustness of Boosting(F) to the ICA factors. On the other hand, if factors are more correctly determined by $r_{PC,IC}$, especially the forecasting accuracy of LARS(F) and CA benefit. For the rolling window forecasts, similar patterns can be observed. The wins of Boosting(F) peak at 15 for $r_{AIC,SIC}$ and LARS(X) wins are highest for r_3 and lowest for $r_{PC,IC}$. Again models like LARS(F) benefit when $r_{PC,IC}$ is used at the expense of LARS(X) and Boosting(F).

The impact of dimensionality reduction techniques is isolated with the help of Table 7. For

this table, the wins are counted for all factor selection methodologies and both windows, making the columns add up to 3 * 2 * 52 = 312. The first observation that can be made is that there is little difference between the win distribution of the PCA and KPCA column. For all models, the number of wins under PCA and KPCA differ by at most one.

Table 7: Wins per model for different dimensionality reduction techniques counted over both windows and all three factor selection methodologies

Model	PCA	ICA	SPCA	KPCA	Total
AR(p)	0	0	0	0	0
ARX(p)	3	3	3	3	12
CADL	14	21	12	14	61
CA	23	0	17	23	63
FAAR	17	0	14	18	49
Boosting(X)	64	93	64	64	285
Boosting(F)	16	12	14	15	57
LARS(X)	92	181	108	93	421
LARS(F)	77	2	72	76	227
Mean	6	0	8	6	20

Secondly, the ICA convergence issue can also be seen as under ICA most factor-based models struggle. CA and FAAR wins drop to 0 and LARS(F) wins drop to 2 compared to more than 70 for the other dimensionality reduction techniques. Boosting(F) again shows the biggest resilience and the number of wins decreases only a bit. Nevertheless, the bad performance of the other factor-based models, also makes the Mean forecast worse. Naturally, the non-factor-based models LARS(X), Boosting(X), and CADL are the biggest beneficiaries under ICA.

Thirdly, the SPCA-based results closely resemble the PCA and KPCA results. However, in general, the wins for the factor-based models CA, FAAR, Boosting(F), and LARS(F) remain slightly below their PCA and KPCA values. Those wins are transferred to LARS(X) which attains 108 wins compared to 92 and 93 for the PCA and KPCA results respectively.

Before any conclusions for hypotheses three and four are drawn, the effect of the interaction of factor selection and dimensionality reduction techniques is analyzed jointly. For this, Table 8 shows which dimensionality reduction technique delivered the overall best model for the different factor selection methodologies. The AR, ARX, CADL, LARS(X), and Boosting(X) wins are summarized under the category 'Benchmark model' as they are independent of the dimensionality reduction methodology. Note that the columns do not add up to the 52 countries since

sometimes two models of different categories had the same best $MSFE$ accuracy, getting a win each.

Table 8: Winning model for each country across all dimensionality reduction techniques and benchmark models (AR, ARX, CADL, LARS(X), Boosting(X)) for different ways of selecting the number of factors r

	Expanding window				Rolling window				Both
	r_3	$r_{AIC,SIC}$	$r_{PC,IC}$	Total	r_3	$r_{AIC,SIC}$	$r_{PC,IC}$	Total	Total
Benchmark model	25	24	23	72	26	19	22	67	139
PCA	14	11	10	35	8	12	15	35	70
ICA	1	1	0	2	2	3	3	8	10
SPCA	10	11	14	35	13	11	11	35	70
KPCA	13	5	15	33	10	12	13	35	68

Overall PCA, SPCA, and KPCA based models are the most accurate for one-fifth of the countries. Benchmark models are the lowest $MSFE$ model for roughly the remaining 2/5 of countries and ICA. Between expanding and rolling window results only ICA wins differ, by taking some wins from the benchmark models under the rolling window.

The influence of the factor selection methodology is analyzed further. For the expanding window, the benchmark model and PCA do best when factors are selected according to r_3 and worst when they are selected according to $r_{PC,IC}$. The opposite holds for SPCA. KPCA seems to be the most sensitive to the choice of factor selection methodology, doing worst under $r_{AIC,SIC}$ and best when $r_{PC,IC}$ is used. For PCA and SPCA the pattern reverses for the rolling window. However, also for the rolling window, a drop in benchmark wins is observed when factors are determined by $r_{AIC,SIC}$ and $r_{PC,IC}$. Again KPCA attains the most wins under $r_{PC,IC}$. To summarize, more wins are shifted from the benchmark to factor-based models if the number of factors is determined according to $r_{AIC,SIC}$ or $r_{PC,IC}$, although the magnitude of that shift is somewhat muted. Between $r_{AIC,SIC}$ and $r_{PC,IC}$ only KPCA shows a clear pattern, attaining more wins under the latter.

To relate these patterns to the individual factor-based models, Figure 3 plots the full sample MSFE results for the different dimensionality reduction and factor selection methodologies. The results of Figure 3 are for the expanding window forecasts. The rolling window forecast graphs are presented in Figure A1 in Appendix A, due to their similarity. The lines in the graph show how, for the given dimensionality reduction technique, the average $MSFE$ gain over the AR model changes for the different factor selection methodologies. The ICA lines are not included

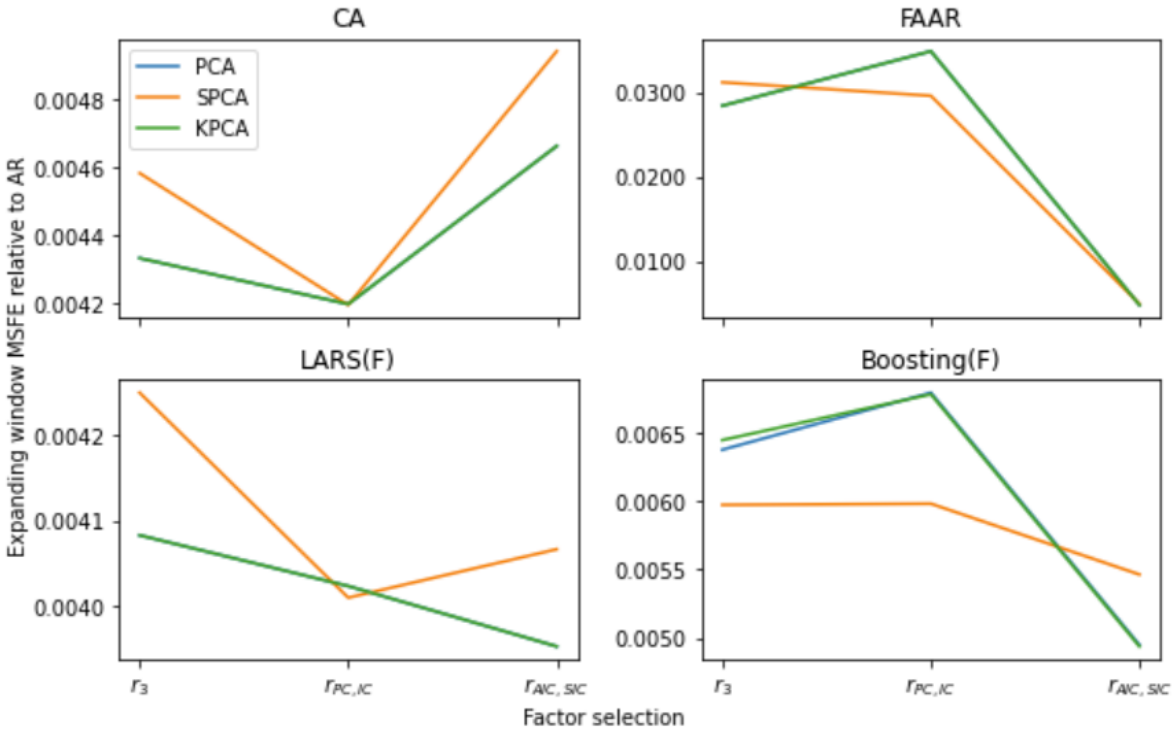

Figure 3: Effect of factor selection methodology on full sample MSFE for expanding window forecasts

since their scales are much bigger.

In general, Figure 3 confirms the patterns detected from Tables 7-8. First of all, PCA and KPCA results closely resemble each other. Therefore, the PCA line is overshadowed by the KPCA for most of the graphs. The CA model attains the best results for PCA, SPCA, and KPCA when the more accurate $r_{PC,IC}$ is used, whereas FAAR and Boosting(F) show the best results for $r_{AIC,SIC}$. The difference between CA and FAAR can be explained by the additional autoregressive terms that influence the FAAR forecasts. From the scale of the y-axis and the number of wins, it is concluded that the additional AR terms in the FAAR worsen forecast accuracy compared to the CA, in line with earlier research of Stock & Watson (2002) and Stock & Watson (2012).

On the other hand, Boosting(F) gives the best results for $r_{AIC,SIC}$ which on average selects more factors than $r_{PC,IC}$. This explains why Boosting(F) wins jumped to 15 for the expanding window in Table 7. Together with the fact that Boosting(X) usually does better than Boosting(F), it seems that Boosting delivers better results if it is applied to more variables and that too much dimensionality reduction harms the forecasting performance. The LARS(F) attains the best results for SPCA when $r_{PC,IC}$ factors are included. However, for PCA and KPCA, $r_{AIC,SIC}$ is preferred which also gives the overall lowest $MSFE$.

Finally, also the Table 7 conclusion that PCA and KPCA are slightly preferred over SPCA is confirmed. LARS(F) and Boosting(F) attain their lowest relative $MSFE$ for PCA and KPCA. For CA and FAAR, the lowest relative $MSFE$ is only dependent on the factor selection methodology since the best model comes from PCA, SPCA, and KPCA. The absolute minimum of these four models is given by LARS(F) when the factors are calculated by PCA or KPCA and selected according to $r_{AIC,SIC}$.

In summary, the factor selection methodology seems to have an impact on forecasting accuracy. Therefore, the findings for hypothesis three are summarized as follows. Selecting factors based on $r_{AIC,SIC}$ and $r_{PC,IC}$ instead of always picking 3 factors r_3 makes factor-based models more accurate and attain more wins at the expense of benchmark models. Under $r_{AIC,SIC}$ this happens mostly due to Boosting(F), which prefers more factors. The more sensitive CA and LARS(F) benefit the most from $r_{PC,IC}$ factors, although LARS(F) delivers the best results for PCA and KPCA together with $r_{AIC,SIC}$. Nevertheless, the changes are somewhat small as LARS(X), LARS(F), and Boosting(X) take the top three places with a big gap to the fourth place for all factor selection methodologies. These are the same models identified as the best models for the simulated datasets and also the somewhat limited impact of $r_{AIC,SIC}$ and $r_{PC,IC}$ is in line with the simulation results.

On the other hand, no clear winner between PCA, ICA, SPCA, and KPCA can be determined. As seen in Table 7 and 8 and Figure 3 PCA, SPCA, and KPCA yield almost identical results, with a slight preference for PCA and KPCA. ICA-based models do the worst, but this is due to a convergence issue of the proposed 'Fast ICA' algorithm. Therefore, it is concluded that KPCA is as good as PCA and slightly preferred over SPCA in this dataset. This answers research question four, on how KPCA based models compare to PCA, ICA, and SPCA-based models. Interestingly, this was expected from the simulation results, as there it was likewise shown that the dimensionality reduction techniques only have very little impact on the best model. This is somewhat in contrast to Kim & Swanson (2018), who find that for h=1 forecasts ICA and SPCA are preferred over PCA. In this paper, PCA does at least as well as SPCA. It might be the case that KPCA won the countries that would have gone to SPCA else, which would put the results of this paper and Kim & Swanson (2018) in line again. However, the similarity of KPCA-based models and PCA-based models, makes it more likely that KPCA took wins away from PCA. Alternatively, this difference in results might simply be due to a different set of models in this paper. As was shown, different models prefer different factor selection and dimensionality reduction techniques. But also the different datasets can potentially explain these somewhat contrasting results.

5.2.4 Additional Results

From an economic perspective, it is also interesting which countries have the highest explanatory power. For this, the variable selection of the best model, LARS(X) is analyzed. Figure 4 shows how often each country was selected as an explanatory variable for the expanding and rolling window forecasts of each country. More detailed country by country results can be found in Figure A2 in Appendix A. Tanzania was an explanatory variable for 162 different forecasts, followed by Djibouti (132), Eritrea (130), and Angola (129). On the other hand, Zambia (24), Nigeria (20), and Equatorial Guinea (15) were selected as the fewest.

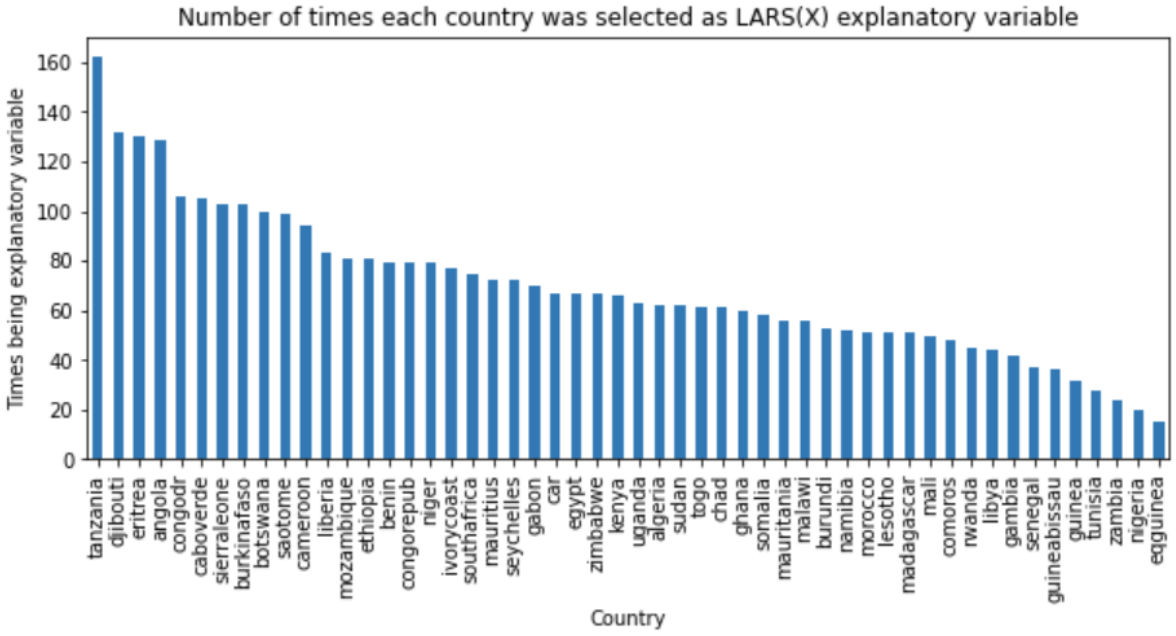

Figure 4: How often each country was selected as explanatory variable in LARS(X) forecasts for rolling and expanding window

Tanzania's high count comes from the forecast of its GDP growth, close countries such as Zambia and Namibia, a cluster of western states including Mali, Burkina Faso, and Sierra Leone, and Eritrea in the northeast. However, for other bordering countries such as Kenya, Uganda, and the Democratic Republic of Congo, Tanzania is never chosen. Similarly, Djibouti is mostly a variable for Central African countries such as Tanzania, the Democratic Republic of Congo, Burundi, and Equatorial Guinea, but also Sierra Leone and Cabo Verde to the very west of Africa. Overall, geographic location does not seem to be important, but also GDP and GDP per capita can't explain the links. The low countries like Nigeria are most of the time not even selected for their forecast, which hints towards the bad performance of AR-type models. Furthermore, high GDP countries such as Egypt, South Africa, and Nigeria, which should have a big impact on the overall African economy, are selected relatively seldom.

Overall, similarity in location and GDP does not seem to drive the explanatory power of some countries. This might be due to only looking at the LARS(X) variable selection, or due to not analyzed economic relations that are more important. Alternatively, this can be interpreted as further evidence for the dissimilarity of economic growth in Africa.

6 Conclusion and Discussion

6.1 Conclusion

This research aimed to extend current literature by replicating the findings of Kim & Swanson (2018) on another dataset and by additionally investigating how KPCA and the factor selection methodology impact forecasting accuracy of factor-based models. Accurate economic forecasts are critical to the policy decisions of central banks and governments all around the world. Analyzing what affects the accuracy of such forecasts, therefore, explains the social relevance of this paper. To answer the four research hypotheses the GDP growth of 52 African countries was forecasted by ten different models under 24 different settings.

The first hypothesis, if factor-based models outperform the forecasting accuracy of AR-type models, can be answered with a clear yes. Across all specifications and in terms of $MSFE$ as well as wins, factor-based models delivered better results than their autoregressive counterparts. These results are in line with an extensive list of past literature such as Stock & Watson (2002), Stock & Watson (2012), Kim & Swanson (2013), and Kim & Swanson (2018). Including autoregressive terms even proved as a disadvantage, as described by Stock & Watson (2002) and Stock & Watson (2012) for the FAAR, which in general performed worse than the CA.

The second hypothesis asked if combining boosting and least angle regression with factor models improves the forecasting performance. Overall LARS(X) delivered the best forecasts in terms of $MSFE$ across all specifications, and Boosting(X) also did better than Boosting(F) under most specifications. However, sometimes LARS(F) is more accurate than LARS(X) and for some specifications even wins more countries. These observations show that there is a benefit in combining boosting and LARS with factorization. Nevertheless, in this dataset arguably LARS(X) is preferred overall, due to the sensitivity of LARS(F) to the right factor methodology. This is in contrast to the findings of Kim & Swanson (2018). But as the research of Bai & Ng (2008), Bai & Ng (2009), Stock & Watson (2012), and even Kim & Swanson (2018) show, different models are preferred for different variables and no single methodology fits all, making general conclusions difficult. This is also emphasized by the country results which showed a big variation in best models across the 52 countries.

The third hypothesis was concerned with the impact of factor selection methodology on forecasting accuracy. Overall, it is shown that $r_{AIC,SIC}$ and $r_{PC,IC}$ improved forecasting performance of factor-based models relative to the fixed r_3. Between $r_{AIC,SIC}$ and $r_{PC,IC}$ the preference changes depending on the model and whether models are evaluated based on $MSFE$ or wins. It is concluded that the better accuracy of $r_{PC,IC}$, as shown by Bai & Ng (2002) and the simulation results, does not necessarily translate to a better forecasting performance.

Finally, to test the fourth hypothesis, KPCA-based models were compared to PCA, ICA, and SPCA-based models. KPCA proved to be quite similar to PCA and is slightly preferred over SPCA. Due to a convergence problem, KPCA is much better than ICA. Unlike Kim & Swanson (2018) SPCA and ICA are not preferred over PCA for the h=1 forecasts. But similar to Cao et al. (2003) a good performance of KPCA is found, although KPCA does not seem to be preferred over PCA. The limitations of these conclusions and starting points for future research are discussed in the next section.

6.2 Limitations and future research

The first limitation of this research is that the proposed 'Fast ICA' algorithm of Hyvärinen & Oja (2000) does not converge when applied to the dataset of this paper. As is evident from the results, ICA still manages to produce meaningful factors occasionally, but on average factor-based models are not doing well with the ICA factors.

The comparison of the different models was also solely based on $MSFE$ and their wins which were also determined by $MSFE$. Due to the number of models, specifications, and countries, it was decided to not further increase dimensionality through additional performance measures. However, for future research, it might be interesting to look at forecasting accuracy differently.

Due to the long runtime and time constraints, the simulation results are only based on one simulated dataset for each DGP. As has been shown by the results of this thesis and prior literature, model preference can vary. To generalize the findings of this paper, research on more simulated data as well as other datasets is necessary.

Finally, this paper only used a subset of the models and specifications proposed by Kim & Swanson (2018) and model hyperparameter were not tuned. Including additional models and finetuning could give a better overview of the best way to construct forecasts from high-dimensional datasets.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433–459.

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

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

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.

Breeden, D. T. (1979). An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics*, *7*(3), 265–296.

Cao, L., Chua, K. S., Chong, W., Lee, H., & Gu, Q. (2003). A comparison of pca, kpca and ica for dimensionality reduction in support vector machine. *Neurocomputing*, *55*(1-2), 321–336.

Donoho, D. L., et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, *1*(2000), 32.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *Annals of statistics*, *32*(2), 407–499.

Fama, E. F. (1992). The cross-section of expected stock returns. *The Journal of Finance*, *47*(2).

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*, 3–56.

Foundation, P. S. (2000–20021). *Python, version 3.8.3.* `https://www.python.org/`.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. doi: 10.1038/s41586-020-2649-2

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Hoffmann, H. (2007). Kernel pca for novelty detection. *Pattern recognition*, *40*(3), 863–874.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science Engineering*, *9*(3), 90-95. doi: 10.1109/MCSE.2007.55

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, *13*(4-5), 411–430.

Kim, H. H., & Swanson, N. R. (2013). Mining big data using parsimonious factor and shrinkage methods. *Available at SSRN 2294110*.

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

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Machine learning proceedings 1992* (pp. 249–256). Elsevier.

Mitra, P., Murthy, C., & Pal, S. K. (2002). Unsupervised feature selection using feature similarity. *IEEE transactions on pattern analysis and machine intelligence*, *24*(3), 301–312.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(85), 2825-2830. Retrieved from `http://jmlr.org/papers/v12/pedregosa11a.html`

Ross, S. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, *13*(3), 341–360.

Schölkopf, B., Smola, A., & Müller, K.-R. (1997). Kernel principal component analysis. In *International conference on artificial neural networks* (pp. 583–588).

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th python in science conference.*

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, *19*(3), 425–442.

Stock, J. H., & Watson, M. W. (1998). Diffusion indexes. *NBER working paper*(w6702).

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

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of forecasting*, *23*(6), 405–430.

Stock, J. H., & Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, *30*(4), 481–493.

Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, *10*(66-71), 13.

Wall, M. E., Rechtsteiner, A., & Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis* (pp. 91–109). Springer.

Wes McKinney. (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (p. 56 - 61). doi: 10.25080/Majora-92bf1922-00a

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

Appendix A - Additional Tables and Figures

Table A1: Descriptive statistics of yearly log growth rates for 52 African countries 1961 - 2015 (T=55)

Country	Mean	Std	Minimum	Median	Max	Skew	Kurtosis	JB
algeria	3.518	7.200	-21.940	3.730	29.491	0.006	6.367	24.566
angola	3.418	7.022	-28.369	3.343	20.376	-1.252	7.428	56.070
benin	3.513	3.039	-5.024	3.826	9.531	-0.920	1.154	14.725
botswana	7.932	5.509	-8.013	7.232	23.428	0.456	1.922	4.319
burkinafaso	4.213	3.047	-1.816	4.210	10.436	-0.068	-0.934	33.570
burundi	2.464	5.602	-14.734	3.440	19.310	-0.150	1.973	2.477
caboverde	5.561	4.517	-2.327	5.164	17.563	0.472	-0.326	25.893
cameroon	3.481	5.441	-11.541	4.114	19.885	-0.161	2.027	2.276
car	0.832	7.408	-45.728	1.980	9.075	-4.762	29.316	1697.066
chad	3.374	8.165	-24.080	2.176	28.968	-0.170	3.205	0.341
comoros	2.993	3.231	-5.551	2.762	12.663	0.108	1.971	2.395
congodr	1.242	6.079	-14.503	1.390	19.227	-0.220	0.849	10.445
congorepub	4.160	5.268	-9.431	3.730	21.188	0.338	2.186	2.425
djibouti	2.208	3.215	-6.828	2.664	6.859	-0.998	0.537	21.785
egypt	4.882	2.662	0.598	4.593	13.628	0.976	1.557	12.760
eqguinea	10.168	19.416	-45.887	8.801	91.629	1.083	5.956	29.099
eritrea	3.888	7.364	-19.237	4.974	19.227	-1.057	1.589	13.993
ethiopia	3.119	6.976	-15.082	3.922	13.015	-0.642	-0.252	26.477
gabon	3.888	9.127	-27.444	4.402	33.289	-0.066	4.765	6.790
gambia	3.755	3.240	-4.395	4.018	11.689	-0.183	0.237	16.829
ghana	3.441	4.386	-13.239	4.306	13.103	-1.402	3.484	17.533
guinea	3.259	1.576	-1.511	3.537	6.297	-0.620	0.702	14.781
guineabissau	1.685	6.937	-32.989	2.762	16.721	-2.511	11.756	220.742
ivorycoast	3.752	5.246	-11.653	3.247	16.212	-0.030	0.336	15.381
kenya	4.522	4.346	-8.121	4.306	20.049	0.664	3.725	4.959
lesotho	4.705	5.527	-14.503	3.826	23.428	0.571	4.601	8.377
liberia	0.267	19.103	-71.335	3.247	72.416	-0.522	6.732	32.534
libya	-0.861	19.713	-97.022	0.100	71.540	-1.253	12.715	218.094

Country	Mean	Std	Minimum	Median	Max	Skew	Kurtosis	JB
madagascar	1.822	4.069	-13.582	2.078	9.440	-1.454	3.936	20.222
malawi	4.138	4.867	-10.759	4.593	15.444	-0.445	1.653	5.647
mali	3.485	5.368	-13.582	3.440	18.482	-0.105	1.742	3.523
mauritania	3.795	5.646	-5.235	2.956	24.451	1.233	2.545	13.619
mauritius	5.424	3.887	-10.647	5.259	13.278	-0.876	4.474	11.356
morocco	4.988	3.981	-5.551	4.688	15.358	0.005	0.410	14.538
mozambique	5.227	5.872	-17.079	6.672	23.744	-0.889	4.520	11.860
namibia	3.985	2.614	-1.816	4.497	11.600	-0.176	0.386	15.068
niger	2.383	5.996	-18.633	2.956	12.663	-1.458	3.780	19.751
nigeria	3.760	7.879	-17.079	4.306	29.043	0.267	2.462	1.248
rwanda	3.905	12.163	-69.716	6.110	30.158	-4.121	25.275	1222.244
saotome	4.354	6.261	-10.870	3.053	24.216	0.758	1.765	8.292
senegal	2.759	3.526	-6.828	3.247	8.526	-0.880	0.552	19.700
seychelles	4.316	5.980	-8.556	4.784	19.227	0.092	-0.261	23.119
sierraleone	2.364	7.281	-22.941	2.859	23.349	-0.837	4.530	11.141
somalia	1.939	9.166	-30.517	2.567	26.313	-0.463	3.033	1.863
southafrica	3.031	2.370	-2.122	3.150	7.603	-0.346	-0.385	25.869
sudan	3.732	5.170	-6.507	4.306	15.444	-0.107	-0.136	21.409
tanzania	4.360	1.973	0.000	4.402	8.158	-0.146	-0.367	24.750
togo	3.623	5.659	-16.370	3.922	14.410	-0.513	1.873	5.029
tunisia	4.514	3.255	-1.918	4.593	16.297	0.743	2.291	5.872
uganda	5.155	3.879	-7.042	5.921	14.669	-0.797	1.415	10.942
zambia	3.155	4.608	-8.992	3.730	15.358	-0.124	0.206	17.049
zimbabwe	2.681	7.485	-19.480	2.567	20.376	-0.610	1.419	8.643

Table A2: All 10 models used for forecasting

Model	Independent of PCA, ICA, SPCA, KPCA
AR(p)	Yes
ARX(p)	Yes
Comb. ADL	Yes
CR	No
FAAR	No
Boosting(X)	Yes
Boosting(F)	No
LARS(X)	Yes
LARS(F)	No
Mean	No

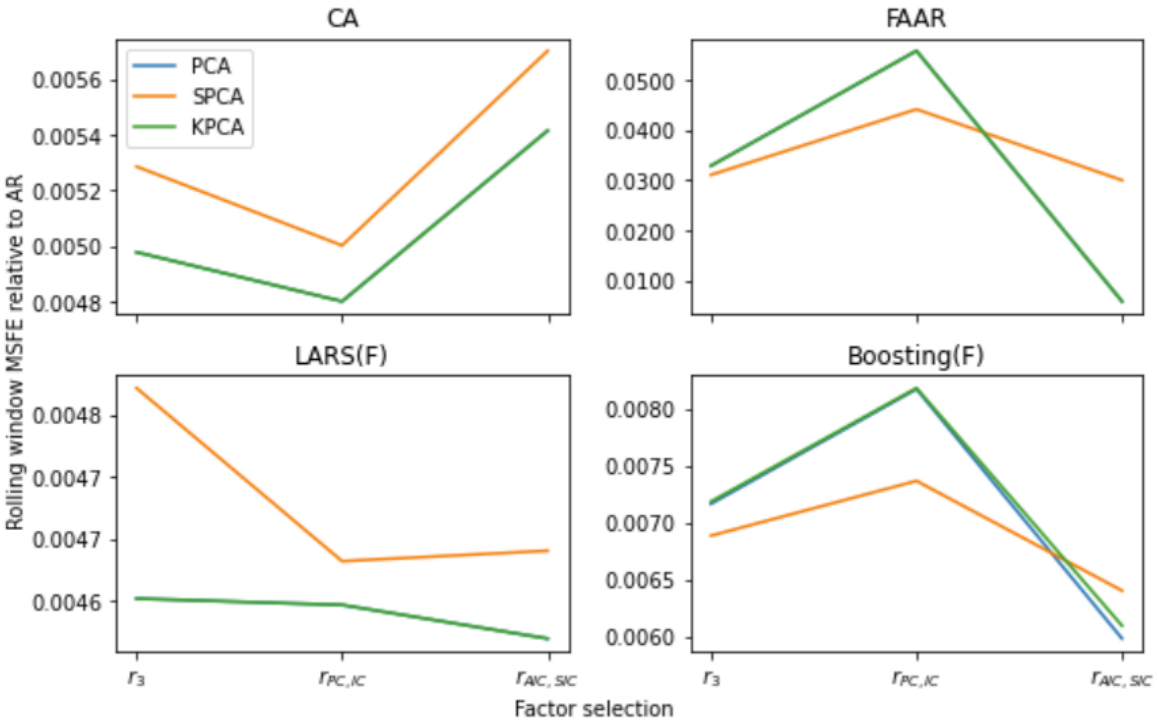

Figure A1: Effect of factor selection methodology on full sample MSFE for rolling window forecasts

Table A3: Best model and accuracy of $r_{PC,IC}$ and $r_{AIC,SIC}$ for rolling window forecasts of the 10 last observation in six simulated datasets for PCA, ICA, SPCA, and KPCA

			$r_{PC,IC}$			$r_{AIC,SIC}$		
	R	Dim. Red	Mean	Accuracy	Best Model	Mean	Accuracy	Best Model
DGP1	1	PCA	1.000	1.000	**LARS(F)**	1.000	1.000	**LARS(F)**
	1	ICA	13.778	0.000	Boosting(F)*	14.333	0.000	LARS(X)
	1	SPCA	1.000	1.000	**CA**	1.111	0.889	**CA**
	1	KPCA	1.000	1.000	**LARS(F)**	1.000	1.000	**LARS(F)**
DGP2	4	PCA	4.000	1.000	CA	5.444	0.667	CA
	4	ICA	14.333	0.000	**FAAR***	17.444	0.000	FAAR
	4	SPCA	4.000	1.000	CA	7.333	0.333	CA
	4	KPCA	4.000	1.000	CA	4.000	1.000	CA
DGP3	8	PCA	7.667	0.667	LARS(X)	8.111	0.889	LARS(X)
	8	ICA	12.667	0.000	**CA***	19.778	0.000	CA
	8	SPCA	8.000	1.000	LARS(X)	11.222	0.111	LARS(F)*
	8	KPCA	7.667	0.667	LARS(X)	8.000	1.000	LARS(X)
DGP4	1	PCA	1.000	1.000	LARS(X)	2.000	0.778	LARS(X)
	1	ICA	12.000	0.111	LARS(X)	15.667	0.000	**Boosting(F)***
	1	SPCA	1.000	1.000	LARS(X)	1.111	0.889	LARS(X)
	1	KPCA	1.000	1.000	LARS(X)	1.000	1.000	LARS(X)
DGP5	4	PCA	4.000	1.000	CADL	4.667	0.667	CADL
	4	ICA	16.667	0.000	CADL	18.667	0.000	**CA***
	4	SPCA	4.000	1.000	CADL	4.222	0.778	CADL
	4	KPCA	4.000	1.000	CADL	4.000	1.000	CADL
DGP6	8	PCA	8.000	1.000	Boosting(X)*	8.111	0.889	CADL
	8	ICA	14.444	0.000	**FAAR***	19.000	0.000	CADL
	8	SPCA	8.000	1.000	Boosting(X)*	10.333	0.444	CADL
	8	KPCA	8.000	1.000	Boosting(X)*	8.000	1.000	CADL

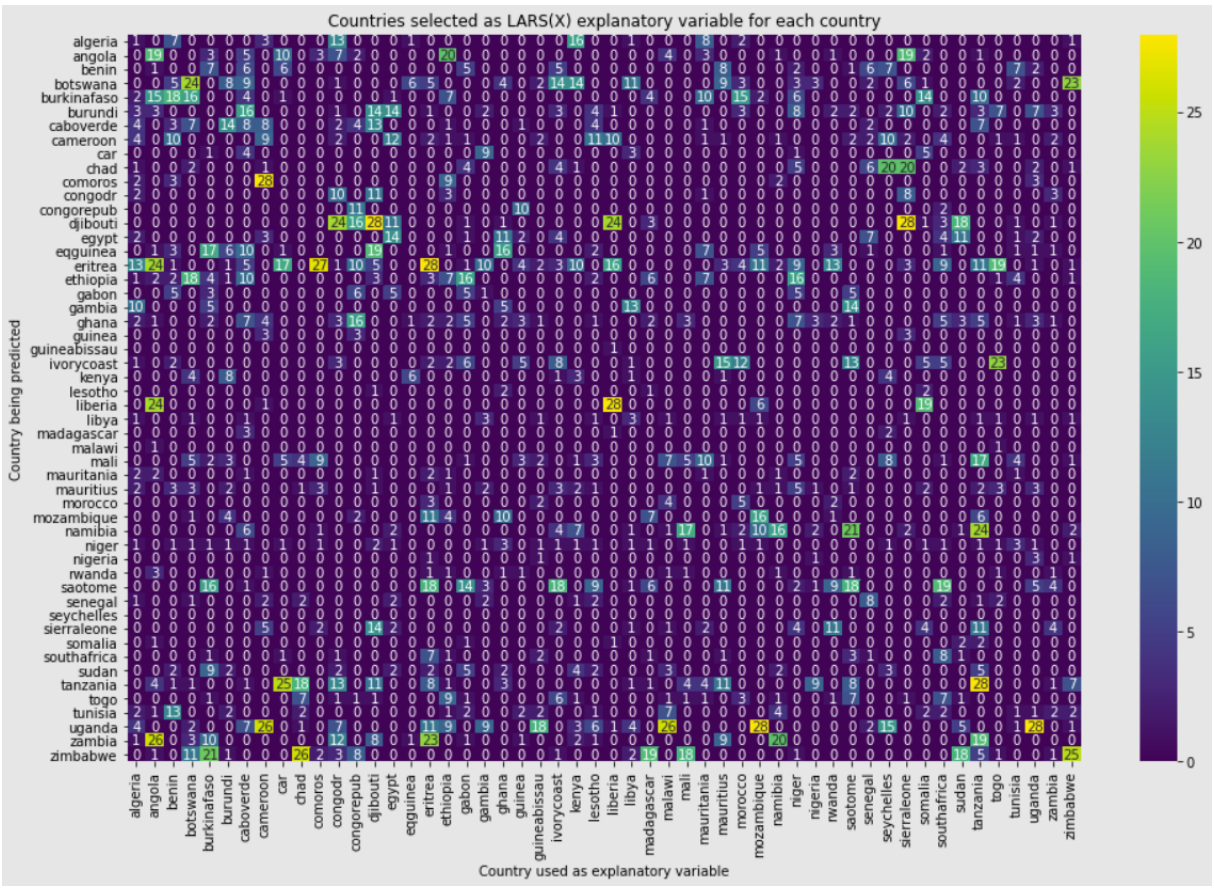

Figure A2: Heatmap showing how often a country was selected as explanatory variable (columns) for the forecasts of all countries (rows)

Appendix B - Code

The entire code of for this thesis was written in Python (Foundation, 2000–20021) and will be uploaded together with the thesis. The following libraries were used: pandas (Wes McKinney, 2010), numpy (Harris et al., 2020), matplotlib (Hunter, 2007), statsmodels (Seabold & Perktold, 2010), and sklearn (Pedregosa et al., 2011).