Can Dimension Reduction and Selection techniques for forecasting beat the simple Average?

Bachelor Thesis Double degree BSc^2 in Econometrics and Economics

By

HaiderUllah Khan

Student No. 457793

Supervisor: Opschoor, P.A.

Second assessor: Tetereva, A

(5-07-2020)

Erasmus School of Economics

Erasmus University Rotterdam

Abstract

In this paper, we try to explore the potential gains of selection and shrinkage of the surveyed forecasts in the ECB Survey of Professional Forecasters. We explore the out-of-sample forecasting accuracy of methods that select, trim or do both and then compare it with the simple average and peLASSO. Our analysis encompasses a wide array of methods including performance-based weighting such as inverse MSPE, statistical combinations based on principal components analysis, and K mean clustering, selection and shrinkage based on regularization techniques such as LASSO, Ridge, Elastic-Net and other egalitarian variants of these methods. Most methods perform better than the simple average and the simple average is beaten by as much as 10 percent. No method except K-mean clustering can beat the peLASSO method. Most of the methods that outperform the simple average perform implicit selection of the best forecaster. We also observe that methods endowed with ex-post information perform equally-well in comparison to methods endowed with ex-ante information.

1 Introduction

Forecasting is a coveted topic in today's world of Economics and Finance. It has become an indispensable part of most companies, be it in terms of financial planning, risk management or operation scheduling. To anticipate the future behavior of the economic variables is quintessential in forecasting future state of the market and financial world. As a result, related organizations can plan effective responsive strategies. An ample amount of literature has been devoted to find the best way to forecast. This paper's aim, though not clear right now, is not centered around constructing forecasting models from inception therefore, we will not explore the topic of forecasting in depth. However, to arrive at the goal of our paper, some basics will be touched upon.

To forecast as accurately as possible, in-depth knowledge of the data generating process, or commonly referred to as the underlying structure of the variable being forecasted, is needed. But most of these are never completely observed. Furthermore, each forecasting model makes careful assumptions. Also, where one forecaster might prefer a non-linear specification, another might choose a linear relationship between the model variables. All these details imply that one forecasting model might be built completely differently from the others despite all giving similar results.

This leads us to the question that in presence of multiple forecasts which type of forecast is the correct one? Bates and Granger (1969) suggest that instead of meticulously identifying the best forecasting model, forecasts from different model can be combined. This prompts the extraction of however much information as could reasonably be expected from these forecasts from different models. A contention in favor can be made that such a combination offers broadening gains. Although, success of such a strategy relies upon the precision of weights utilized (Timmermann (2006)).

This paper centers around forecasts of GDP growth of the EURO Area made by Survey of Professional Forecasters (SPF). In 1999, after the launch of the Euro for the European Union, the European Central Bank (ECB) began SPF to gather information and analyze dynamic Macroeconomic states of EEA. From that point forward, these forecasts from SPF have been utilized as an input for decisions on monetary policy.

However, the forecasted data reported in SPF is generally summarized by using the simple averages of the surveyed forecasts (Genre et al. (2013)). Recent literature is filled

with optimal ways to combine such forecasts (see Clemen (1989), Newbold and Harvey (2002)). However, such techniques are not considered a good choice for the SPF data set due to the unavailability of the methods used by SPF panel members in forecasting Euro Area aggregates. Also, several empirical papers (Genre et al. (2013)) have pointed out that simply averaging the forecasts performs well in comparison with other methods that use estimated combination weights which makes them prone to parameter estimation error. Also, in practice, it is observed that one forecast model outperforms other models in certain periods while other models perform better in other periods. It is difficult to find a forecaster that outperforms all competing forecasters for the entire period. This implies there can be gains from combining different forecasts. Also, due to this time-varying property, forecasts from SPF in this paper are evaluated using different methods on their present and fairly recent performance (using a rolling window). To find forecasts that perform better than individual models and the simple average, two paths can be followed. The first path of combined forecasts can be taken. In a seminal paper, Bates and Granger (1969) construct a combined forecast from a pair of forecasts to achieve overall better predictive performance. This led to combined forecasts being considered as an acceptable alternative to good performing individual forecasts. Succeeding literature (Newbold and Granger (1974); Granger and Newbold (1986), and Yang (2004)) confirmed that indeed, combined forecasts improve forecast accuracy over a single forecast.

In practice, when combining forecasts, we face a key issue of 'dimensionality problem'. This is because a large cross-sectional dimension of SPF forecasts is paired with a comparatively small time series dimension. As combined forecasts mainly rely on estimated combination weights, they are prone to parameter estimation error. This leads us to the second path.

The second path focuses on dimension reduction and selection. The goal of this is to reduce the cross-sectional dimension of the data. Given the dataset and dimensionality problem at hand, second path is the better suited option for the scope of this paper.

Findings from Diebold and Shin (2018) serve as the motivation for the analysis conducted in this paper. Diebold and Shin (2018) introduced partially-egalitarian LASSO (peLASSO) which performs selection in the first stage using LASSO and shrinks the surviving forecasts to equal weights in the second. It is seen that peLASSO performs best among all the regularization techniques and successfully beats the simple average.

As peLASSO encompasses both selection (trim) and Simple average (shrinkage to equal weights), we will use peLASSO (with the simple average in second step) as the benchmark in this paper.

This leads us to the aim of research conducted in this paper which is to forecast quarterly GDP growth of EURO Area by solving the dimensionality problem in selection of forecasts and ultimately, finding selection methods that can beat the simple average and partially egalitarian LASSO in forecasting.

Our examination incorporates an assortment of methods that have been proposed in the existing literature. We incorporate Elastic-Net (Zhu (2005)) with the methods mentioned in Diebold and Shin (2018) as it is also a regularization technique but was skipped in their analysis. We perform i) simple Elastic-Net ii) egalitarian Elastic-Net and iii) partially-egalitarian inspired from Adaptive LASSO (Zhu (2006);Diebold and Shin (2018)). We also perform egalitarian Elastic-Net. For our next method, we consider Integer programming as a selection method (Matsypura et al. (2017)) which provides best forecasts by minimizing over the objective of squared error. This is different than Regularization techniques (including peLASSO) as it doesn't require the estimation of a penalization parameter and it approaches the forecasting problem with a completely different structure.

Motivated from the factor model approach in Hsiao and Wan (2014), we perform Principal Component Analysis (PCA) instead of factor model due to the infeasibility problem in factor models (discussed later in methodology). PCA constructs a parsimonious set of new variables that try to capture most of the variance across all the forecasts whereas regularization techniques select only a few forecasts among the pool. As mentioned above, the goal of combined forecasts is to extract information from the dataset, PCA fulfills this better than peLASSO and integer programming as it includes all the forecasts in analysis when forecasting. This inherent feature of PCA makes it an ideal choice as one of the methods in this paper. PCA like integer programming doesn't require the estimation of a penalization parameter. Selection of K-Mean Clustering (Aiolfi and Timmermann (2006)) and Linear projection (Capistrán and Timmermann (2009)) as methods for conducting analysis in this paper also follow similar reasoning of PCA. However, K-Mean Clustering performed in this paper performs both OLS and the simple average in the second step as opposed to just OLS in existing literature (Aiolfi and Timmermann (2006)). As these

methods mentioned above involve numerous computations and parameter estimation in each step, to counter this we include methods such as Moving Inverse MSPE (Stock and Watson (2001)) and Linear projection in our paper. These methods are computationally simple and possess parsimonious model specification. This makes them less prone to parameter estimation error.

Our paper serves as an extension to the existing literature on dimension reduction as it tries to answer this dilemma in numerous ways by i) removing the worst performers or only basing forecasts over the best performer/performers over a certain time period (for example LASSO, Elastic-Net, K-Mean Clustering, Integer Programming, PElasso) ii) forming a set of new parsimonious variables by trying to capture as much variance across the forecasts as possible (for example Principal Component Analysis and K-mean Clustering) iii) forming combined forecasts based on recent past performance (Moving Inverse MSPE). Our paper is unique in the sense that it compares selection methods from different branches of existing Literature (for example integer programming from Optimization, peLASSO from regularization and PCA from factor model) and tries to evaluate their selection and ultimately, forecasting performance.

Overall, we will use 1) methods that make use of ex-post information. These methods include LASSO, eLASSO, Ridge, eRidge, Elastic Net, pe-LASSO and K-Mean Clustering. 2) methods that only use ex-ante information such as Average-best methods inspired from Diebold and Shin (2018) and, others including Moving Inverse MSPE, Principal Component Analysis, Linear Projection and Integer Programming.

Our main findings include: most methods perform better than simple average and simple average is beaten by as much as 10 percent. No method except K-mean clustering can beat the peLASSO method. We observe that methods endowed with ex-post information perform equally-well in comparison to methods endowed with ex-ante information.

This paper is organized as follows. In Section 2 the methodology used is discussed. Section 3 describes the data and data treatment methods. Section 4 contains the results which are evaluated statistically. Finally, in Section 5, discussion and conclusion are presented.

2 Methodology

2.1 Methods based on ex post information

The machine learning techniques for selection in this section are inspired from Diebold and Shin (2018). In this section, for the tuning parameter in LASSO, eLASSO, Ridge, eRidge, Elastic-Net, eElastic-Net, peLASSO and K-Mean Clustering, we use valuable ex-post information. To find optimal tuning parameters for shrinkage methods, we run a grid search on an equally-spaced grid of [−15,15] which is then exponentiated following Diebold and Shin (2018). For K-mean clustering, the choice of K is also made on silhouette value relying on ex-post information.

2.1.1 Regularization

To forecast a value closer to that of the realized value, this section focuses on regularization methods, a sub-field of machine learning. Due to a large predictor space and a small time series dimension of our dataset, methods relying purely on ordinary least squares (OLS) regression are infeasible (Genre et al.(2013)). OLS selects beta coefficients that minimize residual sum of squares (RSS). This is the deviation of the fitted independent variable from the dependent variable. OLS estimator has the inherent property of being unbiased. This can lead to a dilemma known as the bias-variance trade-off where unbiased-ness comes at a cost of huge variance. This happens when: 1) independent variables (forecasts from forecasters in our case) are highly correlated. 2) There are quite a few independent variables. This becomes more obvious when looking at Equation 1 that shows how variance $\hat{\sigma}^2$ is calculated from the OLS error terms, e.

$$\hat{\sigma}^2 = \frac{e'e}{n-k} \tag{1}$$

where n is the number of observations of the dependent variable, k is the number of dependent variables and e is the residual. When k increases, the variance increases and as k approaches n, the variances nears infinity. This intuitive reasoning implies that presence of bias-variance trade-off can impact forecasting by contaminating the forecasts. To counter this, a bias is introduced which reduces variance. This is known as regularization. James (1961) introduced the first form of shrinkage in literature. This was the first time a bias was added to minimize variance. This was followed by introduction of Ridge

regression (Kennard (1970)) which Myers (1990) summed up perfectly "Ridge regression is one of the more popular, albeit controversial, estimation procedures for combating multicollinearity". Afterwards, Tibshirani (1996) presented Least Absolute Shrinkage and Selection Operator (LASSO) that included ideal properties of feature selection and ridge regression.

We begin the description of regularization methods by starting from the general case. Consider a penalized forecast combining regression, with "parameter budget" c (Diebold and Shin (2018)).

$$\hat{\beta}_{\text{penalized}} = \arg\min_{\beta} \sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it} \right)^2$$
$$\text{s.t.} \sum_{i=1}^{K} |\beta_i|^q \leq c \tag{2}$$

We write this in a Lagrange Multiplier form as follows:

$$\hat{\beta}_{\text{penalized}} = \arg\min_{\beta} \left(\sum_{t=1}^{1} \left(y_t - \sum_{i=1}^{N} \beta_{jt_{it}} \right)^2 + \lambda \sum_{i=1}^{N} |\beta_i|^q \right) \tag{3}$$

Here λ is dependent on c (Diebold and Shin (2018)). λ controls the degree of penalty. Informally stated, amount of shrinkage equals to the choice of λ.

• When λ increases, fewer forecasters are selected (the coefficients of the other forecasters are equal to zero). For a value high enough, no forecasters are selected as coefficients of all forecasters are equal to zero.

• λ is directly proportional to bias.

• λ is inversely proportional to variance. When we include an intercept in the model, it is mostly left unaltered/unchanged. For this reason, we conduct all regularization methods without an intercept.

• When $\lambda = 0$, all forecasters are selected. This implies that no selection but a Bates Granger OLS is performed .

Ridge and eRidge

Since its inception, Ridge regression has become a very common tool for shrinkage in all folds of the Financial world. Ridge regression uses an L2 penalty Kennard (1970). When applying Ridge to the forecasts in our dataset the minimization problem is centered around setting q=2 for equation 4, which shrinks the coefficients toward zero. The

problem is as follows:

$$\hat{\beta}_{\text{Ridge}} = \arg\min_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it} \right)^2 + \lambda \sum_{i=1}^{K} \beta_i^2 \right) \tag{4}$$

As this kind of shrinkage doesn't fall in the scope of our paper because shrinkage technique that this paper revolves around is the simple average. Following Diebold and Shin (2018), we employ a modified ridge regression that shrinks the coefficients toward equality ("egalitarian ridge", or "eRidge"). This happens when the constraint is centered around 1/K as follows:

$$\hat{\beta}_{\text{eridge}} = \arg\min_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta_j f_{it} \right)^2 + \lambda \sum_{i=1}^{K} \left(\beta_i - \frac{1}{K} \right)^2 \right) \tag{5}$$

However, as no selection is performed in Ridge and eRidge, we move on to regularization techniques that fulfill this purpose because selection is the penultimate aim of this paper. Shrinkage methods are only looked at once the selection is made.

Ridge and eRidge are optimized by our function 'Ridge' based on the default 'ridge' function in MATLAB. The default function scales the parameters. We remove this scaling of mean equal to zero from the default function. eRidge is set up based on the instructions listed in Appendix of Diebold and Shin (2018).

LASSO and eLASSO:

Since its advent, LASSO regression (Tibshirani (1996)) has become a major tool for data analysis and theoretical investigations (Knight and Fu (2000) and Meinshausen (2006)). LASSO regression adds an L1 penalty to the error terms of OLS. The minimization problem is as follows:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \left(\sum_{t=1}^{T} \left(y_t - \sum_{i=1}^{K} \beta f_{it} \right)^2 + \lambda \sum_{i=1}^{K} |\beta_i| \right) \tag{6}$$

where T = number of surverys in our dataset, f_i is the forecaster, K is the number of forecasters and k = number of forecasters. Seen before that $q = 2$ results in pure shrinkage (ridge). However, $q = 1$ produces a LASSO estimator that selects and shrinks.This brings us closer to the true model specification.

However, Standard LASSO shrinks the weights of the selected forecasters towards zero.

This type of shrinkage does not fall in the scope of our paper. Therefore, based on Diebold and Shin (2018), we change the specification of LASSO to namely eLASSO:

$$\hat{\beta}_{\text{eLasso}} = \arg\min_{\beta} \left(\sum_{t=1}^{\text{T}} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it} \right)^2 + \lambda \sum_{i=1}^{K} \left| \beta_i - \frac{1}{K} \right| \right) \tag{7}$$

Like LASSO, eLASSO selects and shrinks, but whereas eLASSO shrinks in the right direction, it selects in the wrong direction (Diebold and Shin (2018)). The same problem persists for Ridge and eRidge. Therefore, we now turn to the method that shrinks and selects in the right direction.

LASSO and eLASSO are optimized by default 'lasso' function in MATLAB. The intercept is set to false in this default function. eLASSO is set up based on the instructions listed in Appendix of Diebold and Shin (2018).

Partially-egalitarian LASSO

Following Diebold and Shin (2018), we modify eLASSO such that if performs selection just like LASSO (pure selection) but those selected are shrunk towards equality (this is what eLASSO was already doing). This procedure is named 'partially egalitarian LASSO' and was first introduced by Diebold and Shin (2018). peLASSO falls perfectly under the scope of our paper as it selects (dimension reduction) and then shrinks the survivors to equal weights . This procedure is as following:

$$\hat{\beta}_{\text{peLASSO=}} \arg\min \left(\sum_{f}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it} \right)^2 + \lambda_1 \sum_{i=1}^{K} |\beta_i| + \lambda_2 \sum_{i=1}^{K} \left| \beta_i - \frac{1}{p(\beta)} \right| \right) \tag{8}$$

where $p(\beta)$ is the number of non-zero elements in β. As mentioned in Diebold and Shin (2018) that "The optimization of this one-step objective proves difficult, due to the discontinuity of the objective function at $\beta_i = 0$ " therefore, we imitate the 2-step implementation of Diebold and Shin (2018) as follows:

Step 1 (Select to zero): Using LASSO, select k forecasts from among the full set of K forecasts.

Step 2 (Shrink towards equality): Using standard methods, shrink the combining weights on the k forecasts that survive step 1 toward equality. In step 2, eLASSO, eRidge and Simple average are implemented for shrinkage respectively. This means that we will carry out peLASSO 3 times, each with a different step 2.

Elastic-Net, eElastic-Net and pe-Elastic-Net

There are certain limitations when applying the LASSO method. For example, in a large p, small n case, LASSO selects at most n variables before it saturates. Also, in case of highly correlated variables, LASSO tends to select one variable from this group of correlated variables and ignores the others (Zhu (2005)). As Figure 6 (See Appendix) shows that forecasters in our dataset are highly correlated, this will contaminate the results. To overcome these limitations, the Elastic-net method by Zhu (2005), was introduced.

$$\hat{\beta}_{\text{Elastic-Net=}} \arg\min \left(\sum_{f}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it} \right)^2 + \lambda_1 \sum_{i=1}^{K} |\beta_i| + \lambda \sum_{i=1}^{K} \beta_i^2 \right) \tag{9}$$

The elastic net adds a quadratic part to the penalty ($\|\beta\|^2$) of LASSO's. This penalty when used alone forms ridge regression. The quadratic penalty term makes the minimization problem strongly convex and therefore, it has a unique minimum. The elastic net method includes both LASSO and ridge regression. Each of them is a special case where $\lambda_1 = \lambda, \lambda_2 = 0$ (LASSO) or $\lambda_1 = 0, \lambda_2 = \lambda$ (Ridge). As seen previously q=2 produced shrinkage (Ridge) and q=1 produced selection (LASSO), for both shrinkage and selection (Elastic-Net) $q \in$ (1,2). Based on literature, q is usually chosen midway between LASSO and Ridge i.e. q=1.5. In our paper, we perform a grid-search over (1,2) for selection of optimal q. We also perform egalitarian Elastic-Net for similar reasons as mentioned above for eLASSO and eRidge as follows:

$$\hat{\beta}_{\text{eElastic-Net=}} \arg\min \left(\sum_{f}^{T} \left(y_t - \sum_{i=1}^{K} \beta_i f_{it} \right)^2 + \lambda_1 \sum_{i=1}^{K} \left| \beta_i - \frac{1}{K} \right| + \lambda \sum_{i=1}^{K} (\beta_i^2 - \frac{1}{K}) \right) \tag{10}$$

Inspired from adaptive Elastic-Net in Zhu (2006) and partially-egalitarian methods in Diebold and Shin (2018), we also perform partially-egalitarian Elastic-Net. In the first step, the Elastic-Net is used for selection and the survivors are averaged in the second step.

Elastic-net is implemented in MATLAB using 'lasso' function and setting alpha=0.5 in this function.

2.1.2 K-Mean Clustering:

K-Mean Clustering constructs k clusters that captures information across all the forecasts whereas regularization techniques select only a few forecasts among the pool. As

mentioned above, the goal of combining forecasts is to extract as much information as possible, K-Mean Clustering fulfills this better than peLASSO as it includes all the forecasts in analysis when forecasting. This inherent feature of K-Mean Clustering makes it an ideal choice as one of the methods for our paper.

A key problem when applying regression analysis to the SPF forecasts is the uneven combination of dimensions i.e relatively large cross sectional dimension of available forecasts to be combined together with the relatively small time series dimension. As a simple solution to this, Aiolfi and Timmermann (2006) suggest the use of clustering techniques, applied on the panel of individual forecasts to identify the group structure of the dataset. Aiolfi and Timmermann (2006) split the forecasts into clusters based on their past performance (judged on MSPE) where cluster 1 represents the lowest MSPE's of the forecasts up to cluster K with the highest MSPE's. For all these K clusters, the mean of all the forecasts in each cluster is taken as a regressor, resulting in K independent variables (regressors) which are then regressed on the realized Value of GDP Growth.

However, before applying this method, a decision on the choice of K is critical, as the cluster formation/splitting of forecasts should be based on some criterion. For our paper, based on Baridam (2012), we select K based on mean silhouette value. In this, the observations are split into clusters by minimizing a distance objective function. Observations of mean squared forecast error from t to t-4 against the squared error at time t are plotted. These observations are split into clusters according to their similarity in forecast quality. The similarity within a group is measured by distance of each point to its respective cluster centroid. The centroid is chosen at random but this is updated for each iteration. To this end, different choices as distance measures can be made for example: squared Euclidian distance, sample correlation related measure and cosine measure. For our paper, we opt for Euclidian distance which is the most commonly used in Financial literature.

The performance of the split is evaluated by a silhouette value, s_i, which measures the fit to a cluster and weighs the within-group fit against the fit of the other clusters.

To find optimal K, we set K=2,..,7 which divides the dataset in K clusters. The forecasts are then split and grouped into their respective groups. These forecasts are then compared to their respective group members and to the forecasters in the other K-1 clusters. A mean of silhouette values for all forecasts is taken and plotted against the value of K.

The K that produces the highest mean silhouette value is chosen as optimal K.

Next, to determine weights, the mean each cluster is used as a regressor in a regression on the realized GDP growth value at t. Therefore, K should not be to large, because a larger K increases estimation error. We will perform this method with and without the intercept, and compare them since the coefficient could compensate for a present bias in the forecasts.

$$y_t = \sum_{k=1}^{K} \beta_k \mu_k + \varepsilon_t \tag{11}$$

where μ_k is the mean of each cluster k. Out of sample forecasts are made by multiplying the coefficients, β_i, with the respective mean, μ_k, of each cluster k for that period. As this paper, primarily focuses on Selection/Dimension Reduction, using a regression to derive weights falls in the scope of shrinkage. Therefore, to determine weights, we also take a simple average of the average value of each cluster k to calculate one year ahead forecasts.

As one of criterion in extracting silhouette value for K is forming a t-4 time series therefore, we start our evaluation from the 9th observation onwards. Observations from t=4 to t=8 are used in evaluation. Until t=23, we employ an expanding window and observations used in evaluation are <20. For t >23, we use a rolling window of width 20.

Due to the random nature of selecting the centroids in beginning of K-Mean Clustering, we perform a simulation 100,200 and 500 times and so on until the lowest RMSE found stabilizes . It could be interesting to see if only the best clusters get weights (worst clusters get a coefficient of zero). This would imply that for an accurate forecast, good forecasters will encompass all the information and the worse forecasters are ignored completely.

For finding the K through silhouette value, evalclusters func in MATLAB is used. For performing K-mean clustering, kmeans func in MATLAB is used with a fixed k which was determined in the previous step.

2.2 Enhancing ex-ante forecasting performance

2.2.1 Principal Component Analysis

PCA constructs a parsimonious set of new variables that try to capture most of the variance across all the forecasts whereas regularization techniques select only a few forecasts among the pool. Therefore, PCA extracts more information from the available forecasters

than peLASSO as it includes all the forecasts in analysis when forecasting. This led to selection of PCA as one of the methods for our paper. Furthermore, PCA does not rely on ex-post information compared to K-mean clustering.

Application of PCA in this paper is based on Stock and Watson (2004) who use PCA to estimate common factors from cross-sectional dataset to produce a combined forecast. This technique revolves around capturing the variation in the dataset by means of orthogonal, uncorrelated factors. Every factor is a linear combination of the forecasts in our dataset. The combinations that maximize variance are the eigenvectors of the covariance matrix, ordered from the largest eigenvalue to the smallest. The Principal components are computed by first multiplying factor loadings with the original dataset of forecasters and then summing them. As we have a limited dataset, this might lead to noise in our estimations. This will contaminate the dataset. To tackle this problem, we use an expanding window starting at $t = 5$. This means that for forecasting the GDP growth for the last period, whole dataset is used to perform PCA. As done throughout the paper, for the sake of comparison, we will also perform PCA with rolling window of 20 observations.

For implementation, we compute first few principal components recursively and then regress them on the realized GDP growth rate. Denoting $P_{1,}, \ldots, P_p$ as the first p principal components, we compute the combined forecast using the weights from Equation 12 as follows:

$$y_t = wP_{1,t} + \ldots. + w_pP_{p,t} + \varepsilon_t \tag{12}$$

where w is respective weight for each principal component ,and $t = t$, $t-1$, $t-2$,...,$t-19$ for the rolling window and $t = t$, $t-1$, $t-2$, ..., 1 for the expanding window. These weights are then used to form 1 year ahead out-of-sample forecasts. For application in this paper, we consider combined forecast's performance (Genre et al. (2013)) for up to four principal components (p = 1, 2, 3, 4) and conclude the optimal number of p based on RMSE of the forecasts. This conclusion based on p is a slight ex-post feature of this method.

We use 'pcacov' function in MATLAB to perform this method.

2.2.2 “Average best" combination

Methods mentioned previously rely on information that is not available ex-ante. In a real world setting these methods might perform poorly. Also, these methods rely on estimation of numerous parameters involving complex computation. We select this simple method as it focuses on constructing forecasts by only using ex-ante information and also, is computationally simple and involves the estimation of comparatively less parameters. This makes it less sensitive to parameter estimation error.

Based on the results of Average best combination in Diebold and Shin (2018), we apply a similar method that select only a few forecasts and average them. This eliminates the need for selection of a tuning parameter. For selection, we consider two approaches :1) an individual approach 2) a portfolio approach.

Individual-based average-best combination

For each window of 20 periods, we select the best N individual forecasters over this window and average their 1-year-ahead forecasts. These combinations require selection of N. As N is not known ex-ante, there is an ex-post element in this method when selecting the optimal N. We perform individual-based average-best with i) fixed N where N=1,2,..,6 and ii) N_{max} (Diebold and Shin (2018)). In N_{max}, we examine the past performance of average-best N for N=1,. . . ,N_{max} every period and select the one with lowest RMSE, and use the forecasters in that N to make a one-year ahead forecast. This method is called as “individual-based average best $\geq N_{max}$" forecast combination.

Portfolio (LASSO)-based average-best combination

The portfolio perspective suggests a similar but also a distinct portfolio-based average-best N strategy i.e. at each time t, we perform LASSO to find the N best forecasters for each time rolling forward and average their forecasts. This is called the “LASSO-based average-best N" forecast combination. Like individual perspective, we also perform LASSO-based average-best $\geq N_{max}$ combinations.

2.2.3 Moving inverse MSPE

The reasons for selection of moving inverse MSPE as for our paper follow similar reasoning as mentioned above for Average-Best. However, in comparison to Average-Best, moving

inverse MSPE requires lesser computations. Also, Average-Best method performs selection and assigns weights based on previous performance whereas moving inverse MSPE does not perform selection but assigns weights based on previous performance. However, as later mentioned when k is increased, poorly performing forecasters are approximately close to zero whereas the best performer is close to 1. This can be called an implicit selection and offers an interesting take on the dimensionality problem.

Stock and Watson (2001) propose a method that ignores the correlation between forecasts and uses relative performance to decide the weights in a combined forecast. The relative performance is measured by Mean Squared Prediction Error (MSPE). The weights derived from MSPE perform well empirically because the computation of the off-diagonal elements of forecast error's covariance matrix is skipped.

However, as it is of more relevance how the forecaster fared over the recent past than the entire time period, MSPE will be calculated with a moving window of length 20, resulting in weights that are based on recent performance. MSPE is calculated as follows:

$$MSPE_{i,t} = \frac{1}{20} \sum_{j=t-19}^{j} \left(y_j - f_{i,j}\right)^2, \qquad i = 1, \ldots, K, \tag{13}$$

where t is the current time period and K is the total number of forecasters. For observations $t < 20$, we employ an expanding window. This means that the first time a weight is estimated using a complete rolling window is at t=20 for forecasts. The weights for each forecast (Stock and Watson(2001)) are calculated as follows:

$$w_{i,t} = \frac{MSPE_{i,t}^{-k}}{\sum_{j=1}^{N} \left(MSPE_{j,t}^{-k}\right)} \tag{14}$$

Following the paper by Marcellino (2004), we calculate inverse mean squared prediction error for each k=1, 2, 5. As k increases, the weights given to the lowest MSPEs (best forecasts) increase drastically. For k=5, the spread of weight between the good and bad performing forecasts is significant.

2.2.4 Linear projection

Capistrán and Timmermann (2009) suggest an alternative to ordinary least squares combination method. This method is focuses on reducing the dimensions of the panel data.

This simple linear projection is relatively parsimonious thus, limiting the impact of parameter estimation errors. This method performs a simple regression of equally weighted forecast (simple average) on realized GDP growth value as follows:

$$y_t = a + \beta \bar{f}_t + \varepsilon_t \tag{15}$$

The coefficient estimate of equally weighted forecast from this regression is used as a weight for the equally weighted forecast for forecast at t+1. We perform this method with and without the intercept (bias adjustment parameter) a.
The intercept is calculated to detect the presence of a bias. This bias is also added to the forecast at $t+1$. If the simple average of forecasts from 23 different forecasters act as a true data generating process for realized GDP growth, the coefficient for equally weighted forecast is expected to be closer to 1 and the intercept closer to 0. 0 for intercept implies absence of bias. For forecasts at $t > 20$, rolling window of full length is used.

2.3 Integer Programming:

Matsypura et al. (2017) suggests the use of integer programming to perform selection before shrinking the selected forecasts to equal weights i.e. some weights are equal to each other and sum to unity while the remaining weights are equal to zero. We denote forecasts as $f = (f_1, f_2, ..., f_3) \in \mathbb{R}^n$. The errors for each forecast become $e = \iota y - f = (e_1 e_2 \dots e_n)'$, where ι is an n-vector of ones. Forecast errors are expected to have finite covariance $\mathbb{E}(ee')= \Sigma$. We introduce a vector of weights $w = (w_1 w_2 \dots w_n)' \in \mathbb{R}^n$ to construct a combination forecast. This results in a combination forecast $f^c = w'f$, with error $e^c = y - f^c$. The resulting variance of this combination forecast is $\mathrm{Var}\,(e^c) = w'\Sigma w$. We minimize this variance with constraint $w'\iota = 1$. This yields the weights, in terms of MSE, of the optimal combination as follows:

$$\boldsymbol{w}^{opt} = \underset{w'\iota=1}{\arg\min}\,(\boldsymbol{w'\Sigma w}) = \left(\boldsymbol{\iota'\Sigma^{-1}\iota}\right)^{-1}\boldsymbol{\Sigma^{-1}\iota} \tag{16}$$

The optimal combination forecast becomes $f^{\mathrm{opt}} = (w^{\mathrm{opt}})' f$. The variance of errors from optimal combination forecast is $\mathrm{Var}(\mathrm{e}^{opt}) = (\iota'\Sigma^{-1}\iota)^{-1}$. We form a vector of equal weights for the combination forecast as follows:

$$w^{avg} = \frac{1}{n}\iota \tag{17}$$

The combination forecast now becomes $f^{avg} = (w^{avg})' f$ with $\text{Var}(e^{avg}) = n^{-2}\iota'\Sigma\iota$. We introduce the variable $\tilde{w} \in \{0,1\}^n$ to formulate a binary integer programming problem. To ensure that objective function remains convex, number of non-zero weights are fixed to a non-negative integer $k \in \{1, 2, \ldots, n\}$. The k elements having weights greater than zero can be written as a subset of W as follows (Matsypura et al. (2017)):

$$\mathcal{W}_k = \left\{ \boldsymbol{w} = \frac{\tilde{\boldsymbol{w}}}{k} | \tilde{\boldsymbol{w}}'\boldsymbol{\iota} = k, \tilde{\boldsymbol{w}} \in \{0,1\}^n \right\} \tag{18}$$

We obtain a weight vector after dividing $\tilde{\boldsymbol{w}}$ by k. In this vector, some elements are equal and these sum to one. The remaining are equal to zero. $\tilde{\boldsymbol{w}}'\boldsymbol{\iota}$ acts as a count of non-zero elements and the restriction $\tilde{\boldsymbol{w}}'\boldsymbol{\iota} \geq 1$ guarantees that indeed k forecasts are exactly selected for forecast combination. To find optimal solution with k equal weights, we minimize as follows

$$\begin{aligned} \min \quad & k^{-2}\tilde{w}'\Sigma\tilde{w} \\ \text{s.t.} \quad & \tilde{w}'\iota \geq 1 \\ & \tilde{\boldsymbol{w}} \in \{0,1\}^n \end{aligned} \tag{19}$$

where $\tilde{\boldsymbol{w}}$ the vector of binary variables and Σ, is the n x n positive-definite co-variance matrix of forecast errors. The k^{-2} term in the objective function ensures that objective is evaluated at w (not at $\tilde{w}$).

The above problem can be solved by a generic integer programming solver because it is a tractable convex optimization problem (Yang (2017)). We solve the problem for all $k \in \{1, 2, \ldots, n\}$ n times and select the best solution. This solution gives the lowest value for the objective function g in Algorithm 1.

Algorithm 1 (Yang (2017)) solves the forecast selection problem, as proposed by Matsypura et al. (2017), where the solution k is only updated if the solution for that k is better than existing solutions. Setting ϵ closer to 0 instead of equal to 0 improves computational time and requires less computational power as for very small improvements in the solution, weights are not updated. However, we set the ϵ=0 as we perform a small number of computations in comparison to Matsypura et al.(2017). We perform a rolling window with width 20 and use weights of selected forecasters for the current period, to forecast one-year ahead. On each iteration, a new covariance matrix is calculated. Matsypura et al. (2017) shows that integer programming problem despite possessing high dimensions can be solved in reasonable time to give optimal solutions. We implement in algorithm

in JAVA using 'CPLEX' and in MATLAB using 'quadprog' optimization.

Algorithm 1: Selection Algorithm for Integer Programming

Result: Optimal w for min Σ

$g^* \leftarrow \inf$

for $k \in 1, 2, ..., 23$ **do**

  min$\tilde{g}(\bar{w}) = k^{-2}\tilde{\boldsymbol{w}}'\boldsymbol{\Sigma}\tilde{\boldsymbol{w}}$ s.t. $\overline{\boldsymbol{w}}'\boldsymbol{\iota} = k, \tilde{\boldsymbol{w}} \in \{0,1\}^n$

  if $g^* - \tilde{g}^* \geq \epsilon$ **then**

    $g^* \leftarrow \tilde{g}^*$;

    $w^* \leftarrow k^{-1}w$;

  end

end

return w^* ;

end procedure

3 Data

The data is extracted from European Central Bank's quarterly Survey of Professional Forecasters. We use quarterly 1-year-ahead forecasts real GDP growth of Euro-area. Capistrán and Timmermann (2009) stress that an empirical obstacle that arises in forecast surveys is that the real-time data set is an unbalanced panel. This happens because of the dynamic behavior of forecasters who enter and exit throughout the life of the survey. In addition, sometimes forecasters don't respond to the survey. This results in missing values.

Following Diebold and Shin (2018), we shortlist the forecasters based on frequency of their responses to the survey (1999Q1–2018Q2). However, even these shortlisted forecasters have missing values and gaps, for reasons mentioned above. For implementation of methods discussed in Section 2, the dataset should be void of these gaps. We use a simple regression approach to fill these gaps and balance the panel (Genre et al. (2013)) as follows:

$$f_{i,t} - \bar{f}_t = \beta_i \left(f_{it-1} - \bar{f}_{t-1} \right) + \varepsilon_{it} \tag{20}$$

where $f_{i,t}$ is the i^{th} forecaster at time t. This is an autoregressive process (AR(1)) where current value of a variable depends on its previous value. For equation 20, this means that difference of each forecaster from the simple average in period t depends on the difference

in period $t-1$. Instead of fixing β, we estimate it recursively over the entire period to preserve the artificial real time nature of our subsequent dataset (Genre et al. (2010)). However, this regression is not feasible for estimating the missing values for the first 5 periods, due to insufficient data. We fill in these missing values by taking an average of all the other shortlisted forecasters for each time period. We repeat this for the first 5 periods.

We start with the 1999Q1, coinciding with the period the SPF began and we end with the 2018Q2 survey to ensure that all of our data is of the final revised form following the similar reasons mentioned in Diebold and Shin (2018). We perform the forecast evaluation similar to Diebold and Shin (2018) using RMSE as the evaluation tool. We use the period 1999Q1–2018Q2 for forecasting. Our growth rate forecasts then are from 1999Q3–2018Q4. We use first five forecasts in our estimation. This makes our actual evaluation period from 2000Q4–2016Q4. We use a 5-year (20-quarter) rolling window (Diebold and Shin (2018)).

Table 9 (See Appendix) shows the complete summary statistics of our dataset with Mean, Median, Minimum, Maximum and Standard deviation of all the 23 shortlisted forecasters along with Real GDP Growth.

4 Results

The DM tests are inspired from Diebold and Mariano (1995) where statistic of each method is calculated against the simple average. We compute DM as mentioned in Harvey, Leybourne, and Newbold (1999). In total, 23 forecasters were shortlisted.

4.1 Regularization:

Table 1 shows that the simple average performs significantly better altogether at forecasting than the worst forecaster. However, the best forecaster on the other hand, performs better than the simple average by 7 percent. Methods that involve pure selection select, on average, a modest number of forecasters (approximately three) excluding Elastic-Net and peELastic-Net as they select 8.46 and 4.06 respectively. Despite LASSO shrinking towards zero weights rather than equal weights, it performs equally well as the simple average. eRidge and eLASSO perform as well as (slightly better but negligible improvement) the simple average. This is because shrinkage towards equal weights ends up being

strong. As a result, eRidge and eLASSO produce a simple average. No improvement in the RMSEs for eElastic-Net is seen. Partially-egalitarian methods offer first notice-abe improvement over the simple average. peLASSO methods have RMSEs lower than the simple average close to 10 Percent. Out-of-sample RMSEs of the peLASSO and peElastic-Net methods are as good as that of the best forecaster. peLASSO with eRidge and peElastic-Net perform better than all the other methods. Using Median instead of simple average does not offer improvement which shows that data is not contaminated with outliers.

Figures 1a. and 1b. show that as λ increases the RMSEs decreases until it reaches its lowest point for both LASSO and Ridge. However, as it is increases further from that point on, the RMSEs increase again for both methods. This is due to selected forecasters converging to 0 for heavy penalization. Simple average is never beaten and LASSO and Ridge follow similar pattern.

Figures 1c. and 1d. show that as λ increases the RMSEs decreases until it reaches its lowest point for both LASSO and Ridge. Further increase in λ has no effect on RMSEs and they stay constant for both methods. eRidge and eLASSO are nearly indistinguish-able despite their penalization limits being different. Simple Average is nearly beaten.

Figure 1e. shows for peLASSO (with simple average in second step so only 1 penalty parameter) that as λ increases, the RMSE declines slowly until it reaches it's lowest point. At this point, peLASSO beats the simple average convincingly. Further increase in λ increases the RMSE. Figure 1f. and 1g. also show that all peLASSO methods beat simple average and perform well overall. However, all peLASSO methods are endowed with ex-post information. Determination of optimal λ without such information might result in peLASSO methods performing poorly in comparison to the simple average.

Figure 2 shows that minimizing RMSE over q and λ for Elastic-Net and peElastic-Net has mixed results. For Elastic-Net and eElastic-Net, the lowest RMSEs are for q=2 which translates into Elastic-Net and eElastic-Net being purely Ridge. This implies that these 2 methods base heavily on shrinkage. For peElastic-Net, lowest RMSE is seen q=1.1, which is closer to a pure LASSO penalty. This implies that peElastic-net performs LASSO in the first step for selection and simple average for simple average making it a peLASSO (with simple average). This might be true as the RMSEs for both methods are the same.

Table 1: Forecast RMSEs based on ex post optimal λ*

Regulaization Group	RMSE	λ*	#	DM	p-val
Ridge	1.51	3.21	23	-1.18	0.12
LASSO	1.51	0.38	2.57	0.05	0.48
Elastic-Net	1.52	5.66	8.46	-1.05	0.15
eRidge	1.50	25.6 (max)	23	-1.06	0.15
eLASSO	1.51	3.60	23	0.93	0.18
eElastic-Net	1.52	5.66	23	-0.99	0.16
peLASSO(LASSO, Average)	1.43	0.59	2.33	-0.99	0.16
peLASSO(LASSO, eRidge)	1.41	(0.59, max)	2.32	1.11	0.14
peLASSO(LASSO, eLASSO)	1.42	(0.59, 4.19)	2.32	0.99	0.16
peElasticNet	1.41	0.24	4.06	1.86	0.04
Comparisons	RMSE	λ*	#	DM	p-val
Best	1.42	N/A	1	0.75	0.23
90%	1.47	N/A	1	1.04	0.14
Median	1.56	N/A	1	-0.31	0.38
10%	1.70	N/A	1	-1.93	0.03
Worst	1.82	N/A	1	-2.57	0.00
Simple Average	1.52	N/A	23		

4.2 Average Best

Individual-based average-best forecast combination

Table 2 shows that for average best N, the optimum value for N is $\in$ [2,3]. Moreover, the method performs really well. It performs significantly better than the worst forecaster. It also beats the simple average and performs about as good as the best forecaster for $N \in$ [2,3]. For Average-Best N_{max} RMSE declines from N_{max}=1 to N_{max}=2 and then it stabilizes. The average number of forecasters selected per iteration is less than 2 except for N_{max}=6, this shows that most of the information is acquired from the best performer/best performers in the rolling window. The results follow the similar trend as in Diebold and Shin (2018).

LASSO-based average-best forecast combination:

Table 3 shows that LASSO Based average-best N combinations perform better than simple average, beating the simple average for predictive performance at optimal values of N. However, compared to the individual method, LASSO based combinations perform relatively worse. This might be due to the presence multicollinearity effect in a portfolio combination. When correlation between variables are high, LASSO arbitrarily selects one and removes the others. Due to higher correlation, as seen in Figure 6 (See Appendix),

Table 2: Individual-based average-best forecast combination

Average-best N	RMSE	#	DM	p-val
N = 1	1.45	1	-1.00	0.16
N = 2	1.44	2	-1.03	0.16
N = 3	1.44	3	-1.01	0.16
N = 4	1.45	4	-1.00	0.16
N = 5	1.45	5	-1.00	0.16
N = 6	1.45	6	-1.00	0.16
Average-Best N_{max}	RMSE	#	DM	p-val
N_{max}= 1	1.45	1.00	-1.00	0.17
N_{max}= 2	1.44	1.52	-1.00	0.16
N_{max} = 3	1.44	1.70	-1.00	0.16
N_{max} = 4	1.44	1.88	-1.00	0.16
N_{max} = 5	1.44	1.91	-1.00	0.16
N_{max} = 6	1.44	2.01	-1.00	0.16
Comparisons	RMSE	#	DM	p-val
Best	1.42	1	0.75	0.22
90%	1.48	1	1.05	0.15
Median	1.56	1	-0.31	0.38
10%	1.70	1	1.93	0.03
Worst	1.82	1	-2.57	0.00
Average	1.52	23		

even the top performers might have been removed when using LASSO resulting in a comparatively poor performance.

4.3 Principal Component Analysis:

Figure 3 and Table 4 show that the number of optimum principal components for PCA performed with an expanding window equals to 3. RMSE calculated from PCA for optimal p is as good as the simple Average. For $p < 2$, PCA is trivial as RMSEs are high. For rolling Window forecasts, the RMSE declines as p increases. For $p = 1$, RMSE is higher than the worst forecaster. For p=4, RMSE from PCA is highly competitive and as good as of the best forecaster. This shows that when performing PCA, only recent data should be used in construction of components. This follows from the fact that GDP growth rate is dynamic and fluctuating. It might be easy to predict a GDP growth rate looking back only a few months. The more years of data is used in forecasting and calculation of principal components, the more contaminated the dataset will get with

Table 3: LASSO-based average-best forecast combination.

Average-best N	RMSE	#	DM	p-val
N = 1	1.56	1	-0.70	0.20
N = 2	1.53	2	-1.0	0.16
N = 3	1.47	3	-1.0	0.16
N = 4	1.47	4	-1.1	0.15
N = 5	1.48	5	-1.0	0.15
N = 6	1.50	6	-1.0	0.14
Average-Best N_{max}	RMSE	#	DM	p-val
N_{max}=1	1.56	1.00	-0.70	0.20
N_{max}=2	1.56	1.82	-0.96	0.17
N_{max}=3	1.56	2.62	-0.96	0.17
N_{max}=4	1.48	2.94	-0.98	0.15
N_{max}=5	1.48	2.94	-0.98	0.15
N_{max}=6	1.48	2.94	-1,0	0.14
Comparisons	RMSE	#	DM	p-val
Best	1.42	1	0.75	0.22
90%	1.48	1	1.05	0.15
Median	1.56	1	-0.31	0.38
10%	1.70	1	1.93	0.03
Worst	1.82	1	-2.57	0.00
Average	1.52	23		

Table 4: Forecast RMSEs based on Principal Component Analysis

PCA (Expanding Window)	RMSE	DM	p-val
p=1	1.66	-1.03	0.15
p=2	1.56	-0.99	0.16
p=3	1.52	-1.00	0.16
p=4	1.54	-1.00	0.16
PCA (Rolling Window)	RMSE	DM	p-val
p=1	2.00	-1.04	0.15
p=2	1.65	-1.19	0.12
p=3	1.63	-1.53	0.07
p=4	1.42	-1.36	0.09
Comparisons	RMSE	DM	p-val
Best	1.42	0.75	0.22
90%	1.48	1.05	0.15
Median	1.56	-0.31	0.38
10%	1.70	-1.94	0.03
Worst	1.82	-2.56	0.00
Average	1.52		

Table 5: Forecast RMSEs based on Inverse Moving MSPE

Inverse MSPE	RMSE	DM	p-val
k=1	1.46	-0.99	0.16
k=2	1.45	-0.99	0.16
k=5	1.41	-1.00	0.16
Comparisons	RMSE	DM	p-val
Best	1.42	0.75	0.23
90%	1.47	1.05	0.15
Median	1.56	-0.31	0.38
10%	1.70	-1.93	0.03
Worst	1.82	-2.57	0.00
Average	1.52		

these insignificant past observations which will serve as noise.

Figure 3 shows one interesting result. For $p \in [1,3]$, expanding window has a better performance but as p approaches 4, this reverses in favor of rolling window.

4.4 Moving inverse MSPE:

Table 5 shows the optimal value for moving inverse MSPE is at $k = 5$ which gives a RMSE as good as the best forecaster, beating the simple average with almost 10 percent. Using $k > 1$ creates an increase in the divide between weights where the better forecasts get significantly higher weights. For $k = 5$, the best performing forecast/forecasts get all the weights and negligible weights are given to the rest. This implies that the best forecaster gets weight almost equal to 1 while the remaining get weights almost equal to 0. Therefore, this method performs implicit selection of the best forecaster. This explains why this method performs as good as the best forecaster. Incase, the best forecaster is not dominant, and other forecasters have similar RMSEs, this method will then perform shrinkage instead of implicit selection.

4.5 K-mean Clustering:

Figure 4. shows the results of silhouette value method to determine optimal k. $k = 2$ has the highest mean silhouette value and therefore, is the optimal choice.

Using Least Squares Table 6 shows that without the intercept, K mean clustering for simulation ≥ 500 performs as good as the best forecaster and 7 percent better than the

Table 6: Forecast RMSEs based on K-Mean Clustering

Without Intercept	RMSE	DM	p-val
Simulation=100	1.49	-0.78	0.34
Simulation=200	1.45	-1.01	0.16
Simulation=500	1.41	-1.34	0.11
Simulation=750	1.41	-1.34	0.11
With Intercept	RMSE	DM	p-val
Simulation=100	1.61	-1.17	0.40
Simulation=200	1.58	-1.35	0.18
Simulation=500	1.58	-1.35	0.18
With Simple Average			
Simulation=100	1.40	-1.60	0.90
Simulation=200	1.37	-1.67	0.05
Simulation=500	1.36	-1.68	0.05
Simulation=750	1.36	-1.68	0.05
Comparisons	RMSE	DM	p-val
Best	1.42	0.75	0.22
90%	1.48	1.05	0.15
Median	1.56	-0.31	0.38
10%	1.70	-1.93	0.03
Worst	1.82	-2.56	0.00
Average	1.52		

simple average. With intercept, RMSEs are still better than the worst forecaster and but worse off than the simple average. However, as including an intercept increases the RMSE, this shows the absence of a bias in the combination forecast.

Using Simple Average For number of simulations ≥ 100, K-mean clustering performs even better than the best forecaster. This improvement in performance although negligible, still makes it the first method in this paper to beat the best forecaster. RMSEs derived from this method are the lowest across all the methods used in this paper. Figure 5 helps us to understand better the contents of each cluster k. This shows that the number of forecasters in each cluster are dynamic. This can be reasoned for the better performance of K-mean clustering. K-mean clustering forms two clusters with the first cluster having relatively good forecasting performance while the opposite stands for the second cluster. This shows that for each period, the best and the worst performers change and therefore, also does the composition of each cluster.

Table 7: Forecast RMSEs based on Linear Projection

Linear Projection	RMSE	DM	p-val
with intercept	1.66	-1.11	0.13
without intercept	1.45	-1.00	0.16
Comparisons	RMSE	DM	p-val
Best	1.42	0.75	0.23
90%	1.48	1.05	0.15
Median	1.56	-0.31	0.38
10%	1.70	-1.93	0.03
Worst	1.82	-2.57	0.00
Average	1.51		

Table 8: Forecast RMSEs based on Integer Programming

Integer Programming	RMSE	DM	p-val
Rolling Window	1.42	-1.00	0.16
Expanding Window	1.44	-1.00	0.16
Comparisons	RMSE	DM	p-val
Best	1.42	0.75	0.23
90%	1.48	1.05	0.15
Median	1.56	-0.31	0.38
10%	1.70	-1.93	0.03
Worst	1.82	-2.57	0.00
Average	1.52		

4.6 Linear Projection:

Table 7 shows that Linear Projection with an intercept performs better than the worst forecaster but does not outperform simple average. Without intercept, the RMSE is better than the simple average and closer to the best forecaster. This result implies the absence of bias in the forecasts derived from the simple average of all forecasters. The RMSE computed without intercept is competitive and the model in itself is simple and parsimonious indicating at the efficiency of this model when used in a higher dimension setting. This method concludes in the favor of simple average as an unbiased forecast. However, as this method performs better than the simple average in the absence of the intercept, this implies that just averaging all the forecasters for each time period is not enough to make a forecast. As β differs from one, this shows that something more than simple average is needed to make a good forecast.

4.7 Integer Programming:

Table 8 shows that Integer programming with an expanding window, performs better than the simple average. However, for a rolling window, Integer programming does not only beat the simple average by almost 7 percent but performs as good as the best forecaster. The number of variables selected in integer programming for rolling window per window is approximately 1. This implies that for each iteration, integer programming mostly selects the best forecaster. This can be true as the best forecaster and Integer programming have similar RMSEs. Similar to moving inverse MSPE, integer programming also performs an implicit selection of the best forecaster.

5 Discussion and Conclusion

This paper examines the effects of selection and dimension reduction on one-step ahead forecasts obtained from a panel data comprising of 23 forecasters. The penultimate aim of this paper is to form forecasts that outperform the pe-LASSO and simple average. Simple average has proven to be a very solid benchmark in previous research while pe-LASSO remains one of the best forecast combining methods in recent literature.

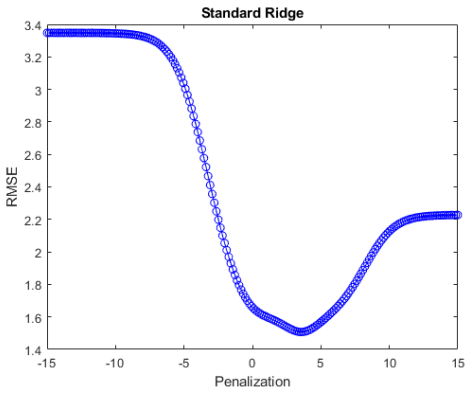
Several conclusions can be drawn from this research. The results from Linear Projection and K-mean clustering show that the simple average of the forecasts from Survey of Professional forecasters does not have a bias. Furthermore, Linear Projection shows that the forecasts made by the simple average do not fully capture the realized value. The inverse MSPE based weights perform very well compared to the mean in the statistical analysis. This stresses out the importance of methods that exploit recent performance for forecasting and give zero/negligible weights to not so recent performance. Another noticeable result shows the application of Principal Component Analysis for dimension reduction and forecast combinations. PCA in such a setting needs further exploration as so far, most work regarding PCA has been done in the field of factor analysis.

Our results also show that some methods, despite being built and computed entirely differently, reduce to performing the implicit selection of the best forecaster. This points out to the limitation of our dataset because our dataset has a dominant best performing forecaster. In absence of such dominant forecaster, and all forecasters performing equally well, these methods are expected to perform shrinkage as well as selection.

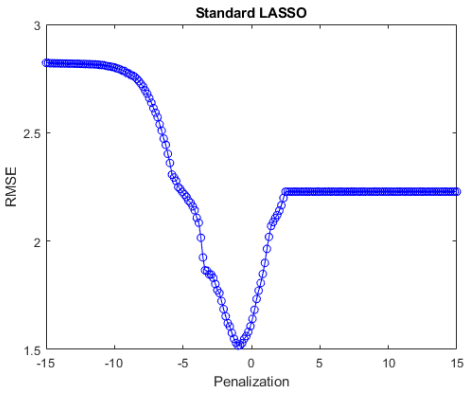
The results from K-mean clustering algorithm are very promising, implying that this method should be a topic for further research. Even though the specification of the K-Mean algorithm is parsimonious and basic, it shows overall better performance than all the methods including both peLASSO and the simple average.

The results from peLASSO involve removing most of the forecasters and simply averaging the survivors. It is seen for peLASSO that the selection penalty needs to be higher as only a few forecasts are combined. In second step, forecast selected for combination are regularized better with regularization techniques. Lastly, the results show that the shrinkage should be extreme, so that the selected forecasts when average give good performing RMSEs. For ex-ante performance, average-best combinations performs well, giving highly competitive MSPEs and beating simple average for optimal N values.

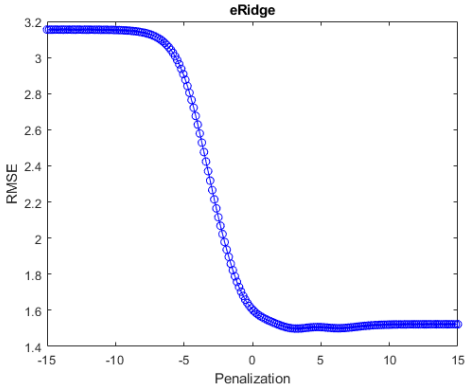
Overall, this paper finds that methods endowed with ex-post information and methods endowed with ex-ante information perform alike. However, most methods perform better than simple average and simple average can indeed, be beaten by as much as 10 percent. No method except K-mean clustering can beat the peLASSO method. This paper points to a topic for further research as the best forecaster remained unbeaten by all except one method. Also, the forecasts from simple average were seen to be missing some information when fitted with the realized values. This implies maybe a shift is needed from the simple average as the most used evaluation technique to only using a parsimonious model based on best forecaster/forecasters.



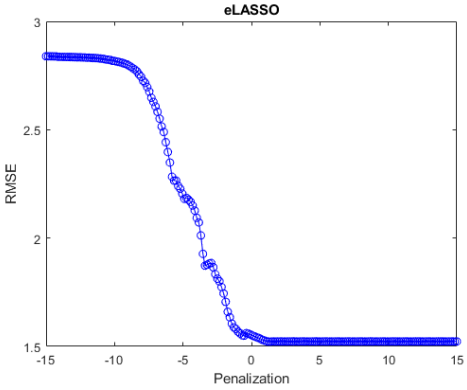
(a) Ridge



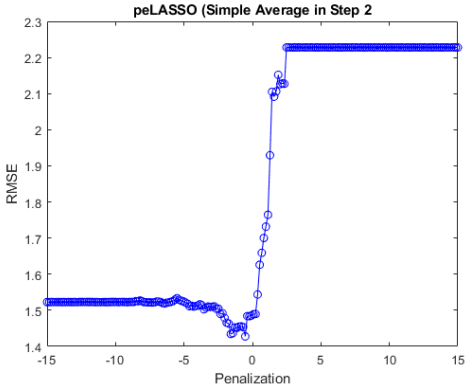
(b) LASSO



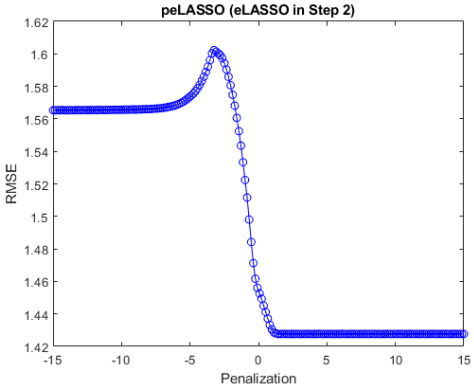
(c) eRidge



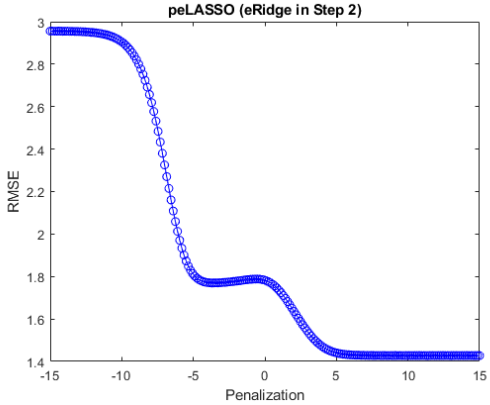
(d) eLASSO



(e) peLASSO (Step 2 Simple Average)

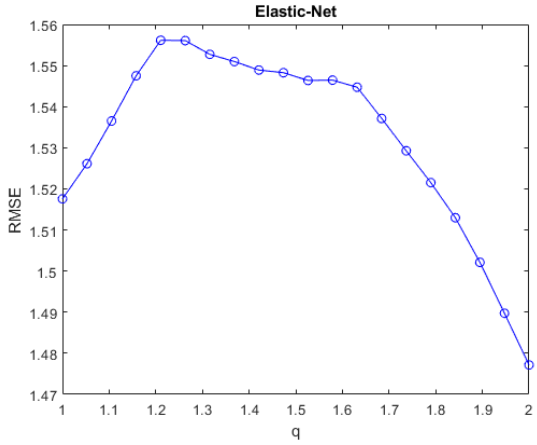


(f) peLASSO (Step 2. with eLASSO)

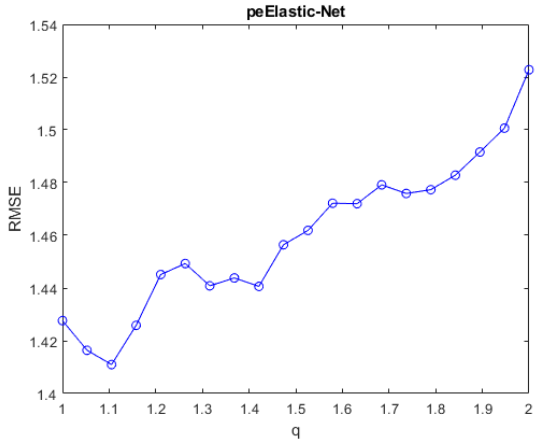


(g) peLASSO (Step 2 eRidge)

Figure 1: RMSE as a function of λ for various forecast combination methods



(a) Elastic-Net



(b) peElastic-Net

Figure 2: RMSE as a function of q for various forecast combination methods

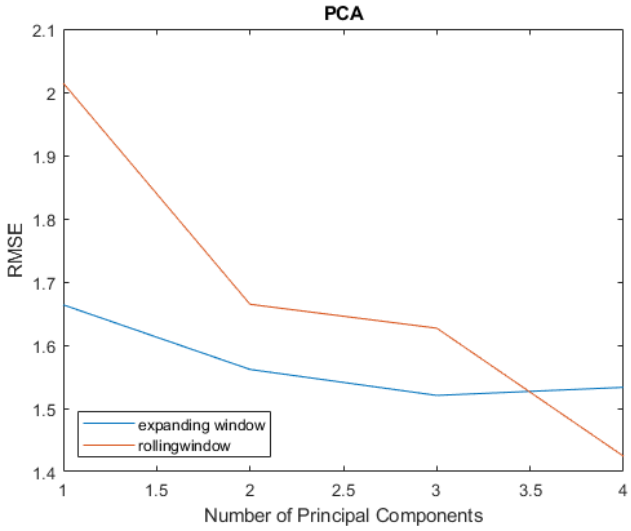


Figure 3: Principal Components with their respective RMSEs

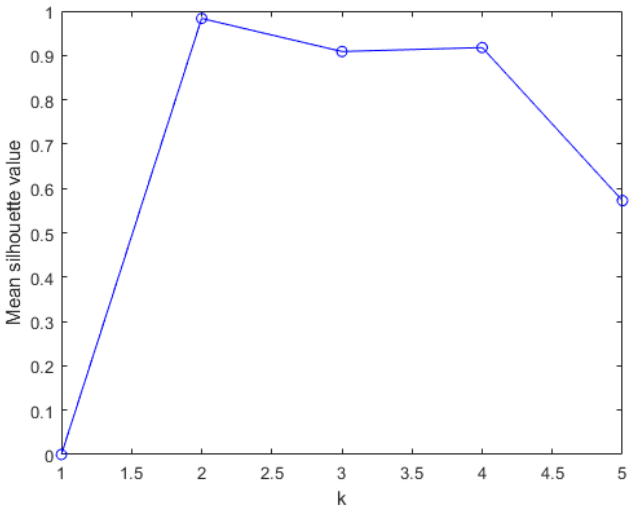


Figure 4: Graph of the value of K against the mean silhouette value. Note: A higher mean silhouette value implies a better overall cluster fit

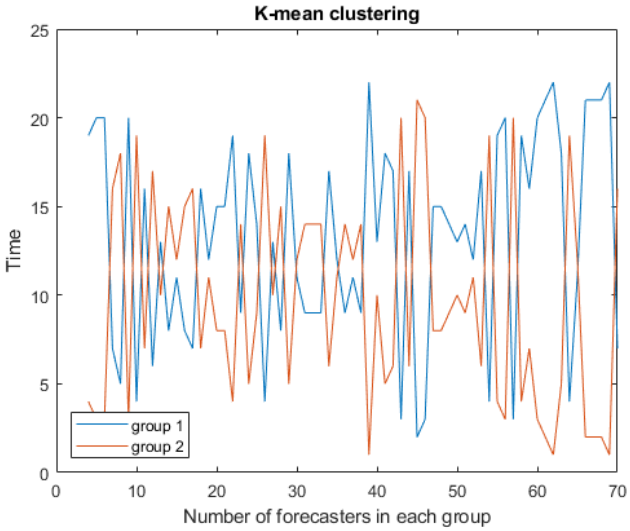


Figure 5: Shows the number of forecasters in each cluster for each quarter

6 References

Aiolfi., M., Timmermann, A., 2006. Persistence in forecasting performance and condi tional combination strategies. *Journal of Econometrics* 135, 31-53

Aiolfi, M., Capistrán, C. and A. Timmermann (2010) "Forecast Combinations", Unpub lished manuscript.

Baridam., B.B., 2012. More work on K -Means Clustering Algorithm: The Dimensionality Problem. *International Journal of Computer Applications* 44(2):23-30

Bates, J., Granger, C. (1969). The combination of forecasts. *Operations Research Quarterly*, 20, 451–468.

Bradley Efron, Trevor Hastie, I. J. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499

Capistrán, C., Timmermann, A. (2009). Forecast combination with entry and exit of experts. *Journal of Business Economic Statistics*, 27(4), 428–440.

Clemen, R. (1989). Combining forecasts: A review and annotated bibliography (with discussion). *International Journal of Forecasting*, 5 (4), 559–583.

Diebold, F., Mariano, R. (1995). Comparing predictive accuracy.*Journal of Business Economic Statistics*, 13, 253–365.

Diebold, F., Pauly, P. (1990). The use of prior information in forecast combination. *International Journal of Forecasting*, 6(4), 503–508.

Diebold, F.X., 1991. A Note on Bayesian Forecast Combination Procedures. In A. Westlund and P.Hackl (eds.) Economic Structural Change: Analysis and Forecasting, *Springer-Verlag*, 225-232.

Diebold, F.X,Shin Michul. (2018) *International Journal of Forecasting* Volume 35, Issue 4, October–December 2019, Pages 1679-1691

Genre, V., Kenny, G., Meyler, A., Timmermann, A. (2013). Combining expert fore casts: Can anything beat the simple average? International *Journal of Forecasting,* 29(1), 108–121.

Granger C. W. J. and P. Newbold (1986) Forecasting Economic Time Series, 2nd Edition, London, Academic Press

Granger, C., Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3, 197–204.

Harvey, D., Leybourne, S., Newbold, P. (1999). Testing the equality of prediction mean squared errors. *International Journal of Forecasting,* 13, 281–291.

Hsiao, C., Wan, S., 2014. Is there an Optimal Forecast Combination? *Journal of Econometrics*, 178, 294-309.

James, W.; Stein, C. (1961). Estimation with quadratic loss. *Journal of political economy*, 1:361–379

Kennard, A. E. H. . R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):607–636.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*,28(5):1356–1378.

Marcellino, M., 2004. Forecast pooling for short time series of macroeconomic variables. *Oxford Bulletin of Economic and Statistics* 66, 91–112.

Matsypura, D., Thompson, R., and Vasnev, A. (2017). Optimal selection of expert forecasts with integer programming.

Nicolai Meinshausen, P. B. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Myers, R. H. (1990). Classical and modern regression with applications

Stock, J., Watson, M. (1999). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In R. Engle, H. White (Eds.), Cointegration, causality, and forecasting. *Oxford University Press.*

Newbold P. and C. W. J. Granger (1974) "Experience with forecasting univariate time series and the combination of forecasts", *Journal of the Royal Statistical Society* Series A, 137, 131-46

Stock, J.H., Watson, M., 2001. A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In: Engle, R.F., White, H. (Eds.), Festschrift in Honour of Clive Granger. Cambridge University Press, Cambridge, pp. 1–44

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

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

Timmermann, Allan G.2005, Forecast Combinations. CEPR Discussion Paper No. 5361. Available at SSRN: http://ssrn.com/abstract=878546

Timmermann, A., 2006, Forecast Combinations, Handbook of Economic Forecasting 1, 135-196.

Yang,.C (2017).Forecast Selection for Combination with Integer Programming: a Robustness Check with a Focus on the Use of Different Estimation Windows (Bachelors Thesis). Available at http://thesis.eur.nl

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society.* Series B. Statistical Methodology, 67, 302–320.

7 Appendix

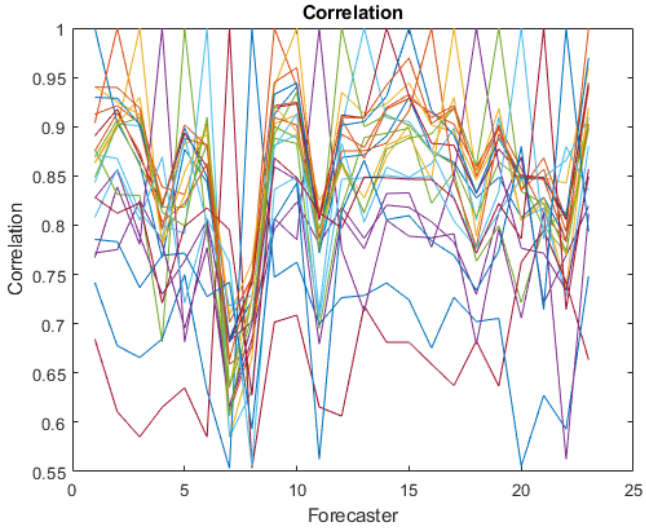


Figure 6: Shows correlation of each shortlisted forecaster with other forecasters

As seen in Figure 6, most of the correlations are above 0.55 and are concentrated mostly around 0.85. This shows that in our dataset a very high correlation is present among forecasters.

Table 9: Summary Statistics of Dataset

	Mean	Median	St. Dev.	Minimum	Maximum
Variables					
Real GDP Growth	1.34	1.80	1.99	-5.70	4.50
SPF Data					
Forecaster 1	1.58	1.60	1.09	-1.50	3.90
Forecaster 2	1.31	1.40	1.29	-3.90	3.90
Forecaster 3	1.41	1.55	1.05	-3.00	3.40
Forecaster 4	1.40	1.70	1.08	-2.80	3.60
Forecaster 5	1.42	1.53	1.01	-2.50	3.60
Forecaster 6	1.26	1.20	1.42	-4.80	3.70
Forecaster 7	1.44	1.40	0.83	-0.30	3.10
Forecaster 8	1.48	1.55	0.84	-1.62	3.40
Forecaster 9	1.48	1.60	1.09	-2.70	3.40
Forecaster 10	1.45	1.69	1.10	-2.00	3.60
Forecaster 1 1	1.26	1.40	1.07	-3.00	3.00
Forecaster 1 2	1.43	1.65	0.97	-2.00	3.60
Forecaster 1 3	1.57	1.79	0.90	-0.90	3.70
Forecaster 1 4	1.32	1.60	1.26	-3.10	3.40
Forecaster 1 5	1.52	1.75	1.03	-2.30	3.40
Forecaster 1 6	1.47	1.71	1.04	-2.50	3.30
Forecaster 1 7	1.47	1.50	0.99	-2.30	3.70
Forecaster 1 8	1.64	1.60	0.78	-0.80	3.00
Forecaster 1 9	1.64	1.77	1.04	-2.50	3.20
Forecaster 20	1.53	1.60	0.96	-2.60	3.00
Forecaster 21	1.49	1.65	1.04	-1.81	3.80
Forecaster 22	1.54	1.67	0.97	-1.20	3.70
Forecaster 23	1.41	1.54	1.03	-2.10	3.30

Forecaster(i) refers to i^{th} forecaster in our SPF dataset of only the most frequent forecasters.

This data has been collected from the European Central Bank (ECB) website