Forecasting Macro-Economic Variables Using Relevant Predictors

Paul Klaassens

July 1, 2013

Abstract

This paper expands on the factor model based forecasts by allowing for more flexibility in the factor model. First a non-linear relationship between the predictors and the factors is allowed by expanding the dataset with the squared predictors. Second a subset of relevant predictors is selected based on their association with the series to be forecast. Hard thresholding, LARS and adaptive lasso techniques are used to select the relevant predictors. Factors are then estimated by applying principal components to these relevant predictors. These relevant predictors change for different dependent variables, forecast horizons and over time. Improvements over the normal factor model forecasts are found for all forecast horizons and variables.

1 Introduction

When forecasting macro-economic variables one often encounters a very large number of possible predictors. Historically academic work has focused on models with only a few variables using model selection to determine the best model. Stock and Watson (2002) have shown that diffusion index forecasts outperform AR, VAR and leading indicator models. Using principal components they estimate factors from a large number of predictors, these factors are the so called diffusion indexes. Then these factors are used to do a linear and direct forecast of y_{t+h}. This way a large portion of the information can be included in the forecast in a simple and compact way.

One of the issues with this method is that the factors that are estimated using principal components try to capture the information from all possible predictors, even though some might not be relevant for forecasting y_{t+h}. As is shown by Boivin and Ng (2006) adding more predictors doesn't always increase forecasting performance. Hence Bai and Ng (2007) suggest doing a prior selection of the relevant predictors before estimating the factors using principal components, or if the relevant predictors are few, use these in a direct forecast. They also allow for a non-linear relationship between the factors and the predictors by including the squared predictors.

The focus of this paper will be on finding the relevant predictors for forecasting y_{t+h} and capturing the information in these predictors in a parsimonious way. There are numerous ways to go about this, Bai and Ng (2007) use hard thresholding to rule out some predictors and least angle regressions to rank them. Mougeot et al. (2012b) introduce the LOL-procedure to reduce the dimensionality by using two thresholding steps. The first step selects the leaders among the predictors, the second step involves performing a linear regression on the leaders and thresholding the result on a certain level. Bair et al. (2006) introduce an algorithm which they call supervised principal components, first a subset of the predictors is selected using univariate regressions after which factors are estimated using principal components.

By adding more flexibility to the factor model of Stock and Watson (2002) this paper aims to improve forecasting accuracy. First the dataset is expanded by adding the squared predictors to allow for a non-linear relationship between the predictors and factors. Secondly, prior to extracting the factors from X_t a selection of the relevant predictors is performed. This selection is achieved in numerous ways. First the hard thresholding technique is used, which is similar to the supervised principal components of Bair et al. (2006). Secondly the adaptive lasso estimator of Zou (2006) is a form of penalized regression which is used to create a subset of relevant predictors. As Efron et al. (2004) have shown, lasso is a special case of Least Angle Regression (LARS) and gives similar results. Here a special case of lasso, the adaptive lasso, is evaluated, which will be compared to LARS. LARS is used to rank the predictors based on when they were selected. By taking the first k predictors a subset of relevant predictors is created. After a subset of relevant predictors is created the factors are estimated by applying the method of principal components to the relevant predictors and the factor forecast is made using the diffusion index framework of Stock and Watson (2002).

2 Background

As model selection has been a popular topic of research there are an abundance of model selection techniques to be found. Some of the most successful and widely known ones include but are not limited to: the ridge estimator of Hoerl and Kennard (1970), the least angle regression of Efron et al. (2004), the Learning of Leaders procedure of Mougeot et al. (2012b) and the lasso estimator of Tibshirani (1996). Recent research has expanded on these selection techniques by adapting and improving them. Zou (2006) proposes a new version of the lasso estimator, the adaptive lasso estimator which has the oracle properties as defined in Fan and Li (2001). Mougeot et al. (2012c) and Mougeot et al. (2012a) both define new selection algorithms which are adaptations of the Learning of Leaders procedure. Also some new estimators have been constructed using a combination of estimators, as the elastic net of Zou and Hastie (2005) is a combination of the ridge and lasso estimators.

In recent years the diffusion indexes of Stock and Watson (2002) have gotten a lot of attention. Instead of selecting only a small set of predictors to use in the model the factors are estimated from a large number of predictors by the method of principal components. These factors are then combined with lags of y and used in a linear forecasting equation to predict y_{t+h}. Stock and Watson (2002) show that these diffusion index forecasts outperform AR, VAR and leading indicator models. Averaging forecasts among multiple factor models has also shown to yield a significant boost in performance as can be seen in van Dongen et al. (2013). Using restricted least squares van Dongen et al. (2013) combine the forecasts of 9 factor models to significantly outperform a single factor model.

One of the issues with factor models is the uncertainty about how much data is needed. As is shown in Boivin and Ng (2006) adding more predictors does not necessarily improve forecasts. If there is too much noise it can be better to remove some predictors before estimating the factors. Boivin and Ng (2006) propose using the correlation between the errors to create a subset of predictors. Bair et al. (2006) introduce a technique they call supervised principal components. A number of predictors are selected based on their relation to y_{t+h} to create a subset of predictors. These subsets are then used to do factor forecasts similar to Stock and Watson (2002), the only difference is that the factors are now extracted from a smaller subset of predictors.

Bai and Ng (2007) expand on these techniques and use least angle regressions, lasso and hard thresholding to determine which predictors to include and estimate factors from. Hard thresholding is very similar to the supervised principal components of Bair et al. (2006). Using univariate regressions of y_{t+h} on the predictors they determine which predictors have significant coefficients and are used for estimating the factors. Bai and Ng (2007) also add the squared predictors to add even more flexibility to the factor structure. Bai and Ng (2007) show significant results using these targeted predictors and outperform the traditional factor model forecasts of Stock and Watson (2002).

As forecasting macro-economic variables using factor models has proven to yield better results than forecasting using traditional methods there is ample reason to do further research on this topic. As is already shown in Bai and Ng (2007) forecasting performance can improve by allowing for more flexibility in the factor structure. By allowing for a non-linear relationship between the factors and predictors and using well-known selection techniques to create a subset of predictors more flexibility is added to the factor model. By adding this flexibility and adjusting the model for different samples, forecast horizons and different y the aim is to yield even better results than the diffusion index forecasts of Stock and Watson (2002).

Forecasting macro-economic variables has been a goal of many econometricians for many years. Expanding on the factor models of Stock and Watson (2002) and following Bai and Ng (2007) the results of this paper can be beneficial to anyone facing the challenge of forecasting a variable with many predictors. As is shown in Bair et al. (2006) the applications of these methods reach further than just economics as they apply the method of supervised principal component to gene expression studies and yield promising results.

3 Data

The same dataset will be used as in van Dongen et al. (2013). This dataset contains 126 variables over a long period of time. The data was then transformed according to the transformation codes which are given by Stock and Watson (2002). After inspection of the data and unit root tests they provided these transformation codes. A complete list of the predictors and their transformations can be found in table 5 in appendix A.2.

After the transformations the data set ranged from March 1960 to September 2009. Any observations exceeding 10 times the interquartile range from the median were marked as harmful outliers and automatically removed. Before constructing factors the data was also standardized.

This paper aims to forecast Personal Income less transfers(PI) and Manufacturing and trade sales(MANU), these are both used in constructing the Index of Coincident Economic Indicators. To perform h-step ahead forecasts the dependent variables are transformed to a h-th difference index. As Personal Income less transfers and Manufacturing and trade sales are both modeled as being first difference in logarithms the following transformation is implemented:

$$y_{t+h}^h = \frac{1200}{h}\ln(\frac{IP_{t+h}}{IP_t}), \quad y_t = 1200\ln(\frac{IP_t}{IP_{t-1}}) \tag{1}$$

4 Method

There are a very large number of possible predictors $X_t^1 = (X_{1,t}, ..., X_{N,t})$ for t = 1,...,T with size N. To allow for a non-linear relationship between the factors and the predictors we can expand the dataset by taking the square of all the transformed predictors and adding them to the dataset. $X_t^{1,2} = (X_{1,t}, ..., X_{N,t}, X_{1,t}^2, ..., X_{N,t}^2)$ for t = 1,...,T with size $2N$.

4.1 Factor model

Let y_{t+h}^h be the variable to be forecast and X_t the set of predictors. W_t is a vector containing a constant and possibly lags of y_t. The following forecasting equation is estimated using data for $t = 1, ..., T-h$:

$$y^h_{t+h} = \alpha' W_t + \beta(L)' \hat{f}_t + \varepsilon_{t+h} \tag{2}$$
$$X_t = \Lambda F_t + \epsilon_t \tag{3}$$

where $\beta(L)$ is a lag polynomial of finite order of at most q, $\hat{f}_t \subset F_t$ and F_t is estimated using principal components. The order of $\beta(L)$ and $\hat{f}_t$ are determined by the BIC. The factor forecast equation is given by:

$$\hat{y}^h_{t+h} = \hat{\alpha}' W_t + \hat{\beta}(L)' \hat{f}_t \tag{4}$$

4.2 Selecting relevant predictors

It is likely that some predictors have more predictive power in forecasting specific y_{t+h}. The factors in the traditional factor model are extracted from the whole dataset X_t. Expanding on this factor model we want to select the relevant predictors beforehand and make a subset $\hat{X}_t \subset X^1_t$ or if the squared predictors are included $\hat{X}_t \subset X^{1,2}_t$. This subset $\hat{X}_t$ can and likely will change for different y, t and h. The factors can then be estimated using principal components:

$$\hat{X}_t = \Lambda F_t + \epsilon_t \tag{5}$$

and equation (2) and (4) can be used to forecast y^h_{t+h}.

4.2.1 Hard thresholding

The hard thresholding approach simply tests if the predictors are significant without considering the correlations between the predictors. This is done by doing an univariate regression for each predictor. Let W_t be a constant and four lags of y_t, execute the following regression:

$$y^h_t = \alpha' W_{t-h} + \phi' X_{i,t-h} + \epsilon_t \quad for\ i = 1, ..., 2N \tag{6}$$

For each of the possible predictors X_i, we get a t-value from the regression, denoted as t_i. For a certain threshold θ we can create a $\hat{X}_t \subset X_t$ by only including X_i if $t_i > \theta$. As there is no clear indication what the exact threshold should be, different values for theta will be tested, $\theta = [2.58, 1.96, 1.65, 1.28]$. These are the two-tailed 1, 5, 10 and 20 percent levels respectively. A 5th forecast is constructed by averaging the forecasts across all thresholds.

The factors can now be estimated using equation (5) and using equation (2) and (4) the factor forecast can be constructed. To evaluate if allowing a non-linear relationship has any performance gains, this selection will be carried out on both X^1_t and $X^{1,2}_t$.

4.2.2 Adaptive Lasso

Another way of reducing the dimension of X_t is by using penalized regressions. Some well-known estimators are the ridge estimator and the LASSO estimator of Tibshirani (1996).

These techniques estimate and select variables simultaneously. The lasso estimator is the solution to:

$$\underset{\beta}{\text{minimize}} \quad RSS + \lambda \sum_{i=1}^{N} |\beta_i| \tag{7}$$

where λ is the tuning parameter which shrinks the coefficients to 0 for large λ and RSS is the sum of squared residuals from regressing y_{t+h}^h on all predictors. It has been shown by Meinshausen and Bühlmann (2006) that under certain conditions the variable selection is consistent. However the lasso yields biased estimates for large coefficients. Fan and Li also suggested that the oracle properties as seen in Fan and Li (2001) do not hold for lasso. Meinshausen and Bühlmann (2006) proved that for the optimal tuning parameter λ, lasso gives inconsistent results.

Zou (2006) shows that the lasso cannot be an oracle procedure and suggests modifying the lasso to allow for different weights for the coefficients. Let w be the weight vector, the modified lasso estimator is given by:

$$\underset{\beta}{\text{minimize}} \quad RSS + \lambda \sum_{i=1}^{N} w_j |\beta_i| \tag{8}$$

If the weights are correctly chosen the modified lasso estimator can have the oracle properties. Suppose $\hat{w} = 1/|\hat{\beta}|^{\gamma}$ and $\hat{beta}$ is a consistent estimator for β, Zou (2006) defines the adaptive lasso estimates by:

$$\hat{\beta}^{*n} = \underset{\beta}{\text{minimize}} \quad RSS + \lambda_n \sum_{i=1}^{N} \hat{w}_j |\beta_i| \tag{9}$$

As the focus of this paper is on selecting relevant predictors to enhance factor model forecasts the estimation of the adaptive lasso will not be beneficiary. Solely the variable selection of the adaptive lasso is of interest. As adaptive lasso sets some of the β coefficients to zero we can form a subset $\hat{X}_t \subset X_t$ only containing the predictors with non-zero coefficients. Factors will be extracted from this subset $\hat{X}_t$ using equation (5) and the factor forecasts proceed as usual using equation (2) and (4).

4.2.3 Least Angle Regressions

The downside of hard thresholding is that it disregards the correlations between the predictors. This could result in a dataset $\hat{X}_t$ with many predictors that have predictive power for y_{t+h} in an univariate setting, but because of the large correlations between the predictors yield no extra information for forecasting. A common way to select predictors is by using forward selection. Based on the correlation with the residual vector predictors are added to the set 1 at a time. Contrary to hard thresholding, forward selection does take the correlation between the predictors into account. Forward selection is too harsh though and often results in too many predictors being eliminated that are correlated with the ones already included. Because of this one might use forward stagewise regression instead, which takes smaller steps.

Let $\hat{\mu}_k$ be the estimate of y with k predictors and $\hat{c} = X'(y - \hat{\mu}_k)$. Now let variable j be the variable for which the absolute correlation is the largest, then the new estimate of y is given by $\hat{\mu}_{k+1} = \hat{\mu}_k + \hat{\gamma} sign(\hat{c}_j) X_j$. As is shown in Efron et al. (2004) forward stagewise regressions are a special case of what is known as Least Angle Regression (LARS). LARS, similarly to forward selection, selects 1 new predictor at each step and adds this variable to the active set. After k steps there will be k variables in the active set and all other variables are excluded. If k is set to the maximum all variables will be added to the set. Since one predictor is added at each step a ranking can be made of all the variables based on when they were added to the active set.

The LARS algorithm starts with $\hat{\mu}_0 = 0$. Define the set of variables with the largest absolute correlations as K and define the current estimate as $\hat{\mu}$.

$$\hat{C} = \max_{j \in K} |\hat{c}_j|, \quad X_k = sign(\hat{c}_j) x_j \quad for\ j \in K$$

Where X_k is the matrix corresponding to K. Now let 1_k be a vector of ones of size K, $G_k = X_k' X_k$ and define:

$$A_k = (1_k' G_K^{-1} 1_K)^{-1/2}, \quad \omega_K = A_K G_K^{-1} 1_K, \quad \mu_K = X_K \omega_K, \quad a_K = X' \mu_K$$

Then the new $\hat{\mu}^* = \hat{\mu} + \hat{\gamma} \mu_K$ where

$$\hat{\gamma} = \min_{j \in K^c}{}^{+} \left(\frac{\hat{C} - \hat{c}_j}{A_K - a_j}, \frac{\hat{C} - \hat{c}_j}{A_K + a_j} \right) \tag{10}$$

As there is no threshold θ or tuning parameter λ to determine the size of the active set, the size of the active set under LARS can be directly chosen by selecting k. To find the optimal number of predictors k^* the Bayesian Information Criterion will be used. Since factors are being estimated from the predictors and the predictors are not used in a direct forecast there are certain flaws with the BIC. Also as shown in van Dongen et al. (2013) and Bai and Ng (2002) the BIC does not perform well for datasets with large N and T. Because of this different predetermined values of k will be used as well, $k = [20, 30, 40, 60, k^*]$.

5 Results

The in-sample period is taken from the start of the data, March 1960 to December 1989. This gives us a large out-of-sample period from January 1990 to September 2009 to evaluate the forecasts and the predictors that are selected. At each t the relevant predictors are selected, new factors are estimated from these relevant predictors and the forecast equation is re-estimated.

5.1 Number of relevant predictors

The first step is selecting the relevant predictors, in table 1 you can find the number of predictors selected averaged over t. The upper part of the table contains the average number of predictors selected for the basic dataset X_t^1 and the bottom half contains the average number of predictors selected for the expanded dataset, $X_t^{1,2}$. These averages are given for both Personal Income less transfers and Manufacturing and trade sales for all forecasts horizons. As described in the method different thresholds for hard thresholding, adaptive Lasso and LARS are used. The $k = [20, 30, 40, 60]$ are omitted in this table since the average number of predictors selected is predetermined by selecting k.

Table 1: *The number of selected predictors averaged over t and rounded to the nearest integer for Personal income(PI) and Manufacturing and trade sales(MANU) for a forecast horizon of* $1, 6, 12$ *and* 24 *months. The averages are shown for hard thresholding with a threshold of* $1\%, 5\%, 10\%$ *and* 20%*, Adaptive lasso with a* γ *of 0.5, 1.0 and 2.0 and LARS with the number of relevant predictors* k^* *selected by the BIC. The top panel contains the results for the normal dataset* X_t^1 *and the bottom panel contains the results for the expanded dataset* $X_t^{1,2}$*.*

X_t^1

Variable	**Horizon**	**Hard thresholding**				**Adaptive lasso**			**LARS**
		1%	5%	10%	20%	$\gamma = 0.5$	$\gamma = 1.0$	$\gamma = 2.0$	k^*
	h = 1	44	59	68	79	9	9	9	9
PI	h = 6	50	62	67	74	12	12	12	14
	h = 12	42	52	59	66	15	15	15	17
	h = 24	34	46	54	64	12	12	12	14
	h = 1	45	59	64	74	21	21	21	23
MANU	h = 6	43	58	68	77	10	10	10	11
	h = 12	47	57	65	74	12	12	12	12
	h = 24	48	56	59	67	17	17	17	22

$X_t^{1,2}$

Variable	**Horizon**	**Hard thresholding**				**Adaptive lasso**			**LARS**
		1%	5%	10%	20%	$\gamma = 0.5$	$\gamma = 1.0$	$\gamma = 2.0$	k^*
	h = 1	63	91	111	134	19	19	19	22
PI	h = 6	96	124	139	158	55	55	55	65
	h = 12	91	119	136	158	80	80	80	137
	h = 24	86	113	127	148	101	104	104	222
	h = 1	66	100	112	133	37	37	37	39
MANU	h = 6	76	108	127	148	47	47	47	50
	h = 12	84	112	129	156	74	74	75	118
	h = 24	101	125	134	153	93	94	94	212

As one would expect hard thresholding selects more variables for $X_t^{1,2}$ and for lower t-value thresholds. As hard thresholding doesn't take the correlation between the selected predictors into account adding the squared predictors was likely to increase the number selected. It's noteworthy to see that this trend isn't only for hard thresholding, as the adaptive lasso and the LARS technique also include more predictors when selecting from $X_t^{1,2}$. We also see that hard thresholding selects on average a lot more predictors than the adaptive lasso and LARS which select similar number of predictors. This is no surprise because of the large correlations between many of the predictors which is disregarded by the hard thresholding approach.

There is also a notable difference between the number of variables selected per dependent variable y_{t+h}. Adaptive Lasso and LARS both select significantly more predictors for MANU than PI when using X_t^1. This difference diminishes when we look at $X_t^{1,2}$, hinting that the squared predictors have predictive power for forecasting PI. Hard thresholding has no significant difference in the number of predictors selected between the different y_{t+h}.

When we look at the different forecast horizons we see that there is no certain trend for X_t^1, however for $X_t^{1,2}$ it is clear that when the forecast horizons changes so do the number of predictors selected. Especially for the longer forecast horizons a large portion of the squared predictors is selected by the adaptive lasso and LARS(k^*). The average number of predictors selected by the LARS(k^*) approach rises very fast for the longer forecast horizons, up to a maximum of 221.51 out of a maximum 252 variables. As noted in van Dongen et al. (2013) and Bai and Ng (2002) the BIC doesn't always perform well for very large datasets with large correlations between the predictors, which could be the cause of this very large number of predictors selected.

5.2 Frequency of predictors

Next we look at how frequent all the predictors are selected. As the predictor selection is performed at each step, which is computational exhausting, it is interesting to see if the relevant predictors change for different t. The results are summarized in table 2 for both dependent variables, forecast horizons and datasets. At each t it is recorded if a predictor is selected or not, by averaging over all t we can get the frequency a predictor is selected ranging from never(0%) until always(100%). Next the selector is placed into 1 of 3 frequency categories, 0-30%, 30-70% and 70-100%, the fraction of the predictors that was placed in each frequency category can be seen in table 2.

A relatively small fraction of the predictors are in the middle frequency category (30-70%). Many of the predictors are either always or never selected, with only a few predictors switching between out and in. This is especially true for X_t^1, when we look at $X_t^{1,2}$ we see that the adaptive lasso and LARS methods for the longer forecast horizons do select different predictors at different t. As the forecast horizon increases it is clear that many of the predictors are more frequently selected. For the 1-month ahead forecast the majority is in the lowest category, but for the 6 and 12-month ahead forecasts the majority is in the middle category. For LARS this shift goes on, for the 24-month ahead forecasts the fractions in the highest category rise to 0.944 for PI and 0.937 for MANU. For adaptive lasso we still see the fraction in the lowest category declining for the 24-month forecast horizon but most of these predictors end up in the middle category instead.

Table 2: *The fraction of predictors that was placed in each frequency category for Personal income(PI) and Manufacturing and trade sales(MANU) for a forecast horizon of* $1, 6, 12$ *and* 24 *months. These are shown for Hard thresholding with a threshold of* 5%, *Adaptive lasso with* $\gamma = 1$ *and LARS with the number of relevant predictors* k^* *selected by the BIC. The top panel contains these results for the normal dataset* X_t^1 *and the bottom panel contains the results for the dataset with the squared predictors included* $X_t^{1,2}$.

X_t^1

Variable	Horizon	Hard thresholding(5%)			Adaptive lasso($\gamma = 1$)			LARS(k^*)		
		0-30%	30-70%	70-100%	0-30%	30-70%	70-100%	0-30%	30-70%	70-100%
	h = 1	0.500	0.087	0.413	0.921	0.024	0.056	0.913	0.032	0.056
PI	h = 6	0.492	0.032	0.476	0.889	0.040	0.071	0.865	0.048	0.087
	h = 12	0.540	0.103	0.357	0.841	0.095	0.064	0.802	0.135	0.064
	h = 24	0.603	0.064	0.333	0.897	0.056	0.048	0.881	0.071	0.048
	h = 1	0.508	0.048	0.444	0.818	0.040	0.143	0.794	0.056	0.151
MANU	h = 6	0.484	0.079	0.437	0.897	0.048	0.056	0.889	0.056	0.056
	h = 12	0.524	0.064	0.413	0.881	0.056	0.064	0.873	0.064	0.064
	h = 24	0.556	0.008	0.437	0.865	0.032	0.103	0.802	0.079	0.119

$X_t^{1,2}$

Variable	Horizon	Hard thresholding(5%)			Adaptive lasso($\gamma = 1$)			LARS(k^*)		
		0-30%	30-70%	70-100%	0-30%	30-70%	70-100%	0-30%	30-70%	70-100%
	h = 1	0.623	0.064	0.314	0.893	0.075	0.032	0.885	0.068	0.048
PI	h = 6	0.480	0.052	0.468	0.742	0.115	0.143	0.702	0.123	0.175
	h = 12	0.476	0.115	0.409	0.627	0.151	0.222	0.131	0.544	0.325
	h = 24	0.520	0.071	0.409	0.464	0.321	0.214	0.036	0.020	0.944
	h = 1	0.575	0.052	0.373	0.837	0.060	0.103	0.829	0.068	0.103
MANU	h = 6	0.524	0.083	0.393	0.742	0.167	0.091	0.738	0.159	0.103
	h = 12	0.520	0.071	0.409	0.631	0.155	0.214	0.552	0.179	0.270
	h = 24	0.492	0.028	0.480	0.536	0.226	0.238	0.028	0.036	0.937

Hard thresholding has a very low fraction of predictors in the middle category all across the board and at closer inspection it is clear that many predictors are either always in or always out. Hence selecting predictors at each t doesn't seem beneficiary as the selected predictors barely change over time. For the adaptive lasso and LARS there is ample reason to select the relevant predictors at each t since the predictors selected fluctuate a lot.

5.3 Forecasting performance

The focus is on improving the factor model by including the squared predictors and selecting the relevant predictors before estimating the factors. Hence a standard factor model as seen in equation (2) and (3) will be used a as a benchmark. The factors are estimated from X_t^1 and a forecast is made using equation (4), that is the BIC selects the number of factors and lags to use for each forecast up to a maximum of 10 factors. The factors are serial, the lags are not. The difference is that the benchmark model extracts factors from X_t^1 and the new factor models select the relevant predictors from X_t^1 or $X_t^{1,2}$ and extracts the factors from the relevant predictors $\hat{X}_t$.

To evaluate the different forecasting techniques the relative mean squared prediction error (RMSPE) is computed as follows:

$$RMSPE = \frac{MSPE(approach)}{MSPE(benchmark)} \tag{11}$$

In table 3 all the RMSPEs are listed. For the adaptive lasso approach the RMSPEs were extremely similar for different γ, hence only the results for $\gamma = 1$ are shown. Under hard thresholding avg is the average forecast taken over all thresholds, and AL is short for adaptive lasso.

Table 3: *Relative mean squared prediction errors for Personal income(PI) and Manufacturing and trade sales(MANU) for forecast horizons of* $1, 6, 12$ *and* 24 *months. Hard thresholding uses four different thresholds:* $1\%, 5\%, 10\%$ *and* 20% *and takes an average (avg) among the thresholds. For adaptive lasso (AL) the parameter* $\gamma = 1$ *and for LARS the number of predictors k is set to* $20, 30, 40, 60$ *or selected by the BIC,* k^**. The top panel contains the results for the normal dataset* X_t^1 *and the bottom panel contains the results for the dataset which includes the squared predictors,* $X_t^{1,2}$*.*

X_t^1

Variable	**Horizon**	**Hard Thresholding**					**AL**	**LARS**				
		1%	5%	10%	20%	avg	$\gamma = 1$	20	30	40	60	k^*
	h = 1	1,00	0,98	0,98	0,98	0,98	1,06	0,98	1,03	1,01	0,98	1,05
PI	h = 6	0,90	0,89	0,90	0,95	0,91	0,90	0,89	0,86	0,93	0,98	0,90
	h = 12	0,92	0,93	0,93	0,98	0,93	0,89	0,81	0,81	0,87	0,92	0,83
	h = 24	1,18	1,05	1,06	1,10	1,08	1,01	0,94	0,95	1,02	1,03	1,00
	h = 1	0,84	0,86	0,87	0,83	0,83	0,83	0,91	0,90	1,04	0,99	0,83
MANU	h = 6	0,92	0,92	0,88	0,86	0,89	0,84	0,86	0,87	0,93	0,96	0,84
	h = 12	0,88	0,92	0,86	0,87	0,87	1,00	0,89	0,92	0,95	1,02	0,98
	h = 24	1,04	0,97	0,97	0,95	0,98	1,20	1,23	1,23	1,09	1,06	1,14

$X_t^{1,2}$

Variable	**Horizon**	**Hard Thresholding**					**AL**	**LARS**				
		1%	5%	10%	20%	avg	$\gamma = 1$	20	30	40	60	k^*
	h = 1	0,97	0,98	1,01	1,00	0,97	1,16	1,08	1,09	1,09	1,04	1,12
PI	h = 6	0,90	0,92	0,96	0,87	0,89	1,27	0,87	0,93	1,31	1,15	1,10
	h = 12	0,91	0,91	0,93	0,94	0,91	0,92	0,83	0,84	0,91	1,05	0,91
	h = 24	1,09	1,08	1,06	1,14	1,08	2,02	1,33	1,73	2,00	1,38	1,09
	h = 1	0,80	0,80	0,81	0,81	0,80	1,04	0,88	1,04	1,03	1,07	1,05
MANU	h = 6	0,95	0,90	0,90	0,89	0,90	1,07	0,94	0,82	0,82	0,91	1,03
	h = 12	0,87	0,84	0,87	0,89	0,86	0,86	1,07	1,01	0,89	0,92	0,81
	h = 24	0,93	0,92	0,89	0,89	0,90	1,48	1,14	1,20	1,11	1,49	1,65

There is a noticeable difference between the forecasting performance for LARS and adaptive lasso when you compare X_t^1 and $X_t^{1,2}$. An example is the 24-month ahead forecast for Personal Income less transfers by adaptive lasso and LARS(40) for $X_t^{1,2}$, the

RMSPEs rise as high as 2.02. In fact, when looking at the LARS and adaptive lasso approach across all thresholds and forecast horizons the forecasts using X_t^1 are better than $X_t^{1,2}$, 23 out of 24 times for PI. The difference is less severe for the MANU forecasts but there is still a significant loss in performance by including the squared predictors.

As the squared predictors are added a large number of them are selected by the LARS and adaptive lasso. As LARS can only select at maximum k predictors this is at the cost of non-squared predictors. Adaptive lasso doesn't have a maximum number of predictors restriction but still drops non-squared predictors because of the correlation with the squared predictors. This seems to hurt the forecasting performance of adaptive lasso and LARS when selecting from $X_t^{1,2}$. When looking at the LARS(k^*), which selects almost all of the predictors the results are not impressive either, but still better than the adaptive lasso and LARS with a set k. This indicates that the dropping of non-squared predictors on behalf of the correlation with squared predictors hurts the performance of the forecasts the most. When we look at hard thresholding we see for PI the results remain quite similar between X_t^1 and $X_t^{1,2}$. However for MANU the forecasts are better when including the squared predictors. Hence adding the squared predictors can have a positive effect on forecast performance if this doesn't result in the dropping of normal predictors which are valuable for forecasting.

When looking at X_t^1 it is clear for LARS(k) that a tight threshold works best. As k increases the RMSPE generally increases as well for both PI and MANU and all forecast horizons. For PI the LARS(k^*) performs average and for MANU the LARS(k^*) outperforms all the other thresholds. The LARS(k^*) selects on average 13,54 predictors for PI and 16,81 predictors for MANU, one can conclude that for this data the optimal number of predictors k^* for LARS is near 16.

Contrary to LARS there is no clear best threshold for hard thresholding. Taking the average across all thresholds sounds promising, as you can hedge yourself against a poorly chosen threshold. In practise though all the forecasts are very similar and the average performs about the same as the sole thresholds. As a result the RMSPEs are very similar for PI for all thresholds and datasets. When looking at MANU there is a minor gain by setting the threshold to 20% and allowing more predictors to be selected. This gain is most noticeable for the longest forecast horizon of 24 months.

Hard thresholding consistently performs better than the benchmark and the other approaches when forecasting MANU, the best forecasts are achieved when using $X_t^{1,2}$ with a threshold of 5,10 or 20%. The average RMSPE across all forecast horizons is 0,87, thus a decrease of 13% in the MSPE. When forecasting PI hard thresholding still does reasonably well for the smaller forecast horizons but for the 24-month ahead forecasts it falls short and performs worse than the benchmark. The best PI forecasts are achieved by the LARS(20) approach, as these are consistently better than the benchmark. The average RMSPE across all forecast horizons is 0,90, thus a 10% decrease in MSPE. When averaging over PI and MANU the all around best approach is still hard thresholding with the predictors selected from $X_t^{1,2}$, this gives us an average decrease in MSPE of 8%.

6 Conclusion

There are significant differences between the optimal number of predictors that the various predictor selection techniques estimate. Hard thresholding selects more predictors from X_t^1 and $X_t^{1,2}$ than the adaptive lasso and LARS(k^*) for most forecast horizons. However when selecting predictors from $X_t^{1,2}$ for a longer forecast horizon the adaptive lasso and LARS(k^*) select a fast growing number of predictors. Hard thresholding also selects a growing number of predictors from $X_t^{1,2}$ for a longer forecast horizon when forecasting MANU but the growth is not as great as for the adaptive lasso and LARS(k^*).

The frequency of which the variables are chosen also differs between the predictor selection techniques. As hard thresholding is quite consistent and selects most of the same predictors at different t the predictors selected by adaptive lasso and LARS fluctuate a lot. Hence this gives good reason to select the relevant predictors at each t, although this is computationally exhausting.

There is a significant decrease in forecasting performance for the adaptive lasso and LARS techniques when adding the squared predictors. When the squared predictors are added to the dataset a large number of these are selected, this is at the cost of some of the non-squared predictors because of the correlation between them. This results in a decrease in forecasting performance as these predictors have more predictive power than their squared counterparts. As hard thresholding does not take the correlation between the predictors into account it doesn't have this problem and can actually benefit from adding the squared predictors.

For each forecast horizon and dependent variable there are numerous forecasts which outperform the traditional factor model of Stock and Watson (2002). There is no clear best approach for selecting the relevant predictors prior to estimating the factors. As hard thresholding with the squared predictors performs best for MANU and LARS(20) performs best for PI. These do give significant gains in forecasting performance of 10% and 13% for PI and MANU respectively.

Therefore it is useful to select the relevant predictors to improve on the factor models of Stock and Watson (2002). The method of hard thresholding has proven to be the most consistent in improving forecasting performance, averaged over PI and MANU this gives an 8% reduction in the MSPE. The Diebold-Mariano test statistics as seen in table 4 in appendix A.1 show that there is significant improvement over the benchmark by the hard thresholding approach with a threshold of 20% for Manufacturing and trade sales for all forecast horizons. However for Personal income the Hard thresholding approach falls short for the longer forecast horizons. Adding the squared predictors didn't show any significant gain in performance for hard thresholding and even worsened the forecasts of the adaptive lasso and LARS.

A Appendix

A.1 Diebold-Mariano test statistic

Let $\hat{y}^h_{bench}$ be the forecasts made by the benchmark model and $\hat{y}^h_{hard}$ be the forecasts made via the hard thresholding technique. The forecast errors and loss differential d are given by:

$$\varepsilon_{bench} = y - \hat{y}^h_{bench}$$
$$\varepsilon_{hard} = y - \hat{y}^h_{hard}$$
$$d = \varepsilon^2_{bench} - \varepsilon^2_{hard}$$

Let $\bar{d}$ be the mean of d and σ^2_d be the variance of d, for a month 1 ahead forecast horizon the Diebold-Mariano test statistic is given by:

$$DM = \frac{\bar{d}}{\sqrt{\frac{1}{T} * \sigma^2_d}}$$

When doing multi-period ahead forecasts there is usually autocorrelation between the forecasts errors. To account for this σ_d is replaced by RV_d:

$$RV_d = \gamma_0 + 2 * \sum_{j=1}^{\infty} \gamma_j, \quad \gamma_j = cov(d_t, d_{t-j})$$

Now $DM \sim N(0,1)$ under the null hypnosis that the forecasting performance of the benchmark model and the hard thresholding technique are equal. We can reject the null hypothesis for a one sided test against the alternative hypothesis that the hard thresholding technique performs significantly better than the benchmark model at the 5% level if $DM > 1.64$. The following table shows the Diebold-Mariano test statistics for hard thresholding using the normal dataset X^1_t.

Table 4: *Diebold-Mariano test statistics comparing the benchmark standard factor model with the hard thresholding technique with* $1\%, 5\%, 10\%$ *and* 20% *thresholds. These are given for Personal income and Manufacturing and trade sales for* $1, 6, 12$ *and* 24 *month forecast horizons.*

X^1_t

Variable	**Horizon**	**Hard thresholding**			
		1%	5%	10%	20%
	h = 1	0.34	1.61	1.37	1.51
PI	h = 6	2.28	2.46	2.30	1.35
	h = 12	2.04	2.43	2.21	0.79
	h = 24	-4.16	-1.20	-1.85	-3.63
	h = 1	1.40	1.30	1.18	1.76
MANU	h = 6	1.55	1.42	2.10	2.42
	h = 12	2.94	2.06	3.45	3.56
	h = 24	-1.69	1.10	1.33	2.04

A.2 List of predictors

A complete list of all the predictors, a brief description and their transformation codes can be found in table 5. The transformation codes are: 1 = no transformation, 2 = first difference, 4 = logarithm, 5 = first difference of logarithms and 6 = second difference of logarithms.

Table 5: *A list of all the 126 predictors used including their transformation codes and a brief description.*

Short name	T-code	Brief description
PI	5	Personal Income (AR, Bil. Chain 2000 $) (TCB)
PI less transfers	5	Personal Income Less Transfer Payments (AR, Bil. Chain 2000 $) (TCB)
Consumption	5	Real Consumption (AC) a0m224/gmdc (a0m224 is from TCB)
M&T sales	5	Manufacturing And Trade Sales (Mil. Chain 1996 $) (TCB)
Retail sales	5	Sales Of Retail Stores (Mil. Chain 2000 $) (TCB)
IP: total	5	Industrial Production Index - Total Index
IP: products	5	Industrial Production Index - Products, Total
IP: final prod	5	Industrial Production Index - Final Products
IP: cons gds	5	Industrial Production Index - Consumer Goods
IP: cons dble	5	Industrial Production Index - Durable Consumer Goods
IP: cons nondble	5	Industrial Production Index - Nondurable Consumer Goods
IP: bus eqpt	5	Industrial Production Index - Business Equipment
IP: matls	5	Industrial Production Index - Materials
IP: dble matls	5	Industrial Production Index - Durable Goods Materials
IP: nondble matls	5	Industrial Production Index - Nondurable Goods Materials
IP: mfg	5	Industrial Production Index - Manufacturing (Sic)
IP: res util	5	Industrial Production Index - Residential Utilities
IP: fuels	5	Industrial Production Index - Fuels
NAPM prodn	1	Napm Production Index (Percent)
Cap util	2	Capacity Utilization (Mfg) (TCB)
Help wanted indx	2	Index Of Help-Wanted Advertising In Newspapers (1967=100;Sa)
Help wanted/emp	2	Employment: Ratio; Help-Wanted Ads:No. Unemployed Clf
Emp CPS total	5	Civilian Labor Force: Employed, Total (Thous.,Sa)
Emp CPS nonag	5	Civilian Labor Force: Employed, Nonagric.Industries (Thous.,Sa)
U: all	2	Unemployment Rate: All Workers, 16 Years & Over (%,Sa)
U: mean duration	2	Unemploy.By Duration: Average(Mean)Duration In Weeks (Sa)
U ¡ 5 wks	5	Unemploy.By Duration: Persons Unempl.Less Than 5 Wks (Thous.,Sa)
U 41760 wks	5	Unemploy.By Duration: Persons Unempl.5 To 14 Wks (Thous.,Sa)
U 15+ wks	5	Unemploy.By Duration: Persons Unempl.15 Wks + (Thous.,Sa)
U 15-26 wks	5	Unemploy.By Duration: Persons Unempl.15 To 26 Wks (Thous.,Sa)
U 27+ wks	5	Unemploy.By Duration: Persons Unempl.27 Wks + (Thous,Sa)
UI claims	5	Average Weekly Initial Claims, Unemploy. Insurance (Thous.) (TCB)
Emp: total	5	Employees On Nonfarm Payrolls: Total Private
Emp: gds prod	5	Employees On Nonfarm Payrolls - Goods-Producing
Emp: mining	5	Employees On Nonfarm Payrolls - Mining
Emp: const	5	Employees On Nonfarm Payrolls - Construction
Emp: mfg	5	Employees On Nonfarm Payrolls - Manufacturing
Emp: dble gds	5	Employees On Nonfarm Payrolls - Durable Goods
Emp: nondbles	5	Employees On Nonfarm Payrolls - Nondurable Goods
Emp: services	5	Employees On Nonfarm Payrolls - Service-Providing
Emp: TTU	5	Employees On Nonfarm Payrolls - Trade, Transportation, And Utilities
Emp: wholesale	5	Employees On Nonfarm Payrolls - Wholesale Trade
Emp: retail	5	Employees On Nonfarm Payrolls - Retail Trade
Emp: FIRE	5	Employees On Nonfarm Payrolls - Financial Activities
Emp: Govt	5	Employees On Nonfarm Payrolls - Government
Emp-hrs nonag	5	Employee Hours In Nonag. Establishments (AR, Bil. Hours) (TCB)
Avg hrs	1	Avg Weekly Hrs of Prod or Nonsup Workers On Private Nonfarm Payrolls - Goods-Producing
Overtime: mfg	2	Avg Weekly Hrs of Prod or Nonsup Workers On Private Nonfarm Payrolls - Mfg Overtime Hours
Avg hrs: mfg	1	Average Weekly Hours, Mfg. (Hours) (TCB)
NAPM empl	1	Napm Employment Index (Percent)
Starts: nonfarm	4	Housing Starts:Nonfarm(1947-58);Total Farm&Nonfarm(1959-)(Thous.,Saar)
Starts: NE	4	Housing Starts:Northeast (Thous.U.)S.A.
Starts: MW	4	Housing Starts:Midwest(Thous.U.)S.A.
Starts: South	4	Housing Starts:South (Thous.U.)S.A.
Starts: West	4	Housing Starts:West (Thous.U.)S.A.

Short name	T-code	Brief description
BP: total	4	Housing Authorized: Total New Priv Housing Units (Thous.,Saar)
BP: NE	4	Houses Authorized By Build. Permits:Northeast(Thou.U.)S.A
BP: MW	4	Houses Authorized By Build. Permits:Midwest(Thou.U.)S.A.
BP: South	4	Houses Authorized By Build. Permits:South(Thou.U.)S.A.
BP: West	4	Houses Authorized By Build. Permits:West(Thou.U.)S.A.
PMI	1	Purchasing Managers' Index (Sa)
NAPM new ordrs	1	Napm New Orders Index (Percent)
NAPM vendor del	1	Napm Vendor Deliveries Index (Percent)
NAPM Invent	1	Napm Inventories Index (Percent)
Orders: cons gds	5	Mfrs' New Orders, Consumer Goods And Materials (Bil. Chain 1982 $) (TCB)
Orders: dble gds	5	Mfrs' New Orders, Durable Goods Industries (Bil. Chain 2000 $) (TCB)
Orders: cap gds	5	Mfrs' New Orders, Nondefense Capital Goods (Mil. Chain 1982 $) (TCB)
Unf orders: dble	5	Mfrs' Unfilled Orders, Durable Goods Indus. (Bil. Chain 2000 $) (TCB)
M&T invent	5	Manufacturing And Trade Inventories (Bil. Chain 2000 $) (TCB)
M&T invent/sales	2	Ratio, Mfg. And Trade Inventories To Sales (Based On Chain 2000 $) (TCB)
M1	6	Money Stock: M1(Curr,Trav.Cks,Dem Dep,Other Ck'able Dep)(Bil$,Sa)
M2	6	Money Stock:M2(M1+O'nite Rps,Euro$,G/P&B/D Mmmfs&Sav&Sm Time Dep(Bil$,Sa)
M2 (real)	5	Money Supply - M2 In 1996 Dollars (Bci)
MB	6	Monetary Base, Adj For Reserve Requirement Changes(Mil$,Sa)
Reserves tot	6	Depository Inst Reserves:Total, Adj For Reserve Req Chgs(Mil$,Sa)
C&I loans	6	Commercial & Industrial Loans Oustanding In 1996 Dollars (Bci)
Cons credit	6	Consumer Credit Outstanding - Nonrevolving(G19)
Inst cred/PI	2	Ratio, Consumer Installment Credit To Personal Income (Pct.) (TCB)
S&P 500	5	S&P's Common Stock Price Index: Composite (1941-43=10)
S&P div yield	2	S&P's Composite Common Stock: Dividend Yield (% Per Annum)
Fed Funds	2	Interest Rate: Federal Funds (Effective) (% Per Annum,Nsa)
Comm paper	2	Cmmercial Paper Rate (AC)
3 mo T-bill	2	Interest Rate: U.S.Treasury Bills,Sec Mkt,3-Mo.(% Per Ann,Nsa)
6 mo T-bill	2	Interest Rate: U.S.Treasury Bills,Sec Mkt,6-Mo.(% Per Ann,Nsa)
1 yr T-bond	2	Interest Rate: U.S.Treasury Const Maturities,1-Yr.(% Per Ann,Nsa)
5 yr T-bond	2	Interest Rate: U.S.Treasury Const Maturities,5-Yr.(% Per Ann,Nsa)
10 yr T-bond	2	Interest Rate: U.S.Treasury Const Maturities,10-Yr.(% Per Ann,Nsa)
Aaa bond	2	Bond Yield: Moody's Aaa Corporate (% Per Annum)
Baa bond	2	Bond Yield: Moody's Baa Corporate (% Per Annum)
CP-FF spread	1	cp90-fyff (AC)
3 mo-FF spread	1	fygm3-fyff (AC)
6 mo-FF spread	1	fygm6-fyff (AC)
1 yr-FF spread	1	fygt1-fyff (AC)
5 yr-FF spread	1	fygt5-fyff (AC)
10 yr-FF spread	1	fygt10-fyff (AC)
Aaa-FF spread	1	fyaaac-fyff (AC)
Baa-FF spread	1	fybaac-fyff (AC)
Ex rate: avg	5	United States;Effective Exchange Rate(Merm)(Index No.)
Ex rate: Switz	5	Foreign Exchange Rate: Switzerland (Swiss Franc Per U.S.$)
Ex rate: Japan	5	Foreign Exchange Rate: Japan (Yen Per U.S.$)
Ex rate: UK	5	Foreign Exchange Rate: United Kingdom (Cents Per Pound)
EX rate: Canada	5	Foreign Exchange Rate: Canada (Canadian $ Per U.S.$)
PPI: fin gds	6	Producer Price Index: Finished Goods (82=100,Sa)
PPI: cons gds	6	Producer Price Index: Finished Consumer Goods (82=100,Sa)
PPI: int matls	6	Producer Price Index:I ntermed Mat.Supplies & Components(82=100,Sa)
PPI: crude matls	6	Producer Price Index: Crude Materials (82=100,Sa)
Spot market price	6	Spot market price index: bls & crb: all commodities(1967=100)
NAPM com price	1	Napm Commodity Prices Index (Percent)
CPI-U: all	6	Cpi-U: All Items (82-84=100,Sa)
CPI-U: apparel	6	Cpi-U: Apparel & Upkeep (82-84=100,Sa)
CPI-U: transp	6	Cpi-U: Transportation (82-84=100,Sa)
CPI-U: medical	6	Cpi-U: Medical Care (82-84=100,Sa)
CPI-U: comm.	6	Cpi-U: Commodities (82-84=100,Sa)
CPI-U: dbles	6	Cpi-U: Durables (82-84=100,Sa)
CPI-U: services	6	Cpi-U: Services (82-84=100,Sa)
CPI-U: ex food	6	Cpi-U: All Items Less Food (82-84=100,Sa)
CPI-U: ex shelter	6	Cpi-U: All Items Less Shelter (82-84=100,Sa)
CPI-U: ex med	6	Cpi-U: All Items Less Medical Care (82-84=100,Sa)
PCE defl	6	Pce, Impl Pr Defl:Pce (1987=100)
PCE defl: dlbes	6	Pce, Impl Pr Defl:Pce; Durables (1987=100)
PCE defl: nondble	6	Pce, Impl Pr Defl:Pce; Nondurables (1996=100)
PCE defl: service	6	Pce, Impl Pr Defl:Pce; Services (1987=100)
AHE: goods	6	Avg Hourly Earnings of Prod or Nonsup Workers On Private Nonfarm Payrolls - Goods-Producing
AHE: const	6	Avg Hourly Earnings of Prod or Nonsup Workers On Private Nonfarm Payrolls - Construction
AHE: mfg	6	Avg Hourly Earnings of Prod or Nonsup Workers On Private Nonfarm Payrolls - Manufacturing
Consumer expect	2	U. Of Mich. Index Of Consumer Expectations(Bcd-83)

References

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

J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317, 2007.

E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

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

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

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

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

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

M. Mougeot, D. Picard, and K. Tribouley. Grouping strategies and thresholding for high dimensional linear models. 2012a.

M. Mougeot, D. Picard, and K. Tribouley. Learning out of leaders. *Journal of the Royal Statistical Society*, 74(3):475–513, 2012b.

M. Mougeot, D. Picard, and K. Tribouley. A new selection method for high-dimensionial instrumental setting : application to the growth rate convergence hypothesis. 2012c.

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

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

T.J. van Dongen, P. Klaassens, J.S. Kop, and L.S. Tijssen. Averaging forecasts across number of factors. 2013.

H. Zou and T Hastie. Regularization and variable selection via te elastic net. *Journal of Royal Statistical Society: Series B*, 67(2):301–320, 2005.

Hi. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.