Linear Supervised Learning Methods for forecasting high-dimensional time series

Samuele Borsini (0001083685)
Mario D'Agostino (0001084078)
Iari Orlandi (0001083849)
Pablo Suárez-Sucunza (0001075377)

November 1, 2023

Abstract

This paper follows De Mol et al. (2008) and addresses the performance of Linear Supervised Learning Methods when forecasting high-dimensional time series. We carry this analysis by predicting CPI using a large macroeconomic time series dataset with four different methods: Principal Component Regression, LASSO, Ridge and Partial Least Squares. We find that the latter three are valid alternatives to PCR, and that their forecasts are highly correlated.

1 Introduction

Many problems in economics require the exploitation of large information. However, large-dimensionality brings along also some estimation issues that recent methods in the literature have tried to fix. The linear supervised learning (LSSL) methods indeed come in handy when OLS, or other least squares estimators, are either overfitting the data or not feasible at all.

LSSL methods can be classified into three categories: subset selection, shrinkage and dimension reduction. In order to handle the dimensionality problem, subset selection methods choose a subset of the available predictors and apply OLS to them; shrinkage methods, Ridge and LASSO, respectively shrink some parameters towards zero and to zero; dimension reduction methods, Principal Component Regression and Partial Least Squares, project the predictors onto a smaller dimensional sub-space and then apply OLS.

Following De Mol et al. (2008), this paper addresses the use of Linear Supervised Learning Methods in forecasting with high-dimensional data with the aim of establishing a connection between them. In particular, we use a standard large macroeconomic dataset to study the forecasting performance of shrinkage and dimension reduction methods, since the subset selection methods are often computationally unfeasible with a large number of predictors.

In doing so, we find that both PLS and shrinkage methods (Ridge and LASSO) offer a valid alternative to PCR, which is used as reference to compare all the results: as in De Mol, the forecasts provided by the methods are highly correlated to those of PCR and are close to the time series of the variable of interest, when considering the optimal level of regularization.

The rest of this paper is organized as follows. Section 2 briefly discusses the main results of the literature. Section 3 reports the characteristics and transformations applied to the dataset used together with the methodology applied in this analysis. Sections 4 to 7 report the theory and the empirical results for the four methods considered. Section 8 concludes.

2 Literature review

De Mol et al. (2008) considers Ridge, LASSO and Principal Component Regression as forecasting methods on a highly collinear panel. They show that, empirically, shrinkage methods' forecasts are highly correlated with principal component forecasts on the dataset exploited by Stock and Watson (2002) that has been used to establish properties of the forecasts of the latter. As a matter of fact, they find that, although Ridge and LASSO (Gaussian and double-exponential prior respectively) rely on different estimation strategies, the two methods produce forecasts which are not only highly correlated but also characterized by similar mean-square errors. Moreover, these forecasts are highly correlated with those produced by principal components, also with similar mean square errors: they do well when PCR does well. Indeed, in De Mol (2008) Ridge and Lasso are shown to be a valid alternative to principal components when it comes to forecasting performance. However, since LASSO performs variable selection, it is signalled to be unstable and very sensitive to minor perturbations of the data. As a matter of fact, variable selection methods are unlikely to provide interpretable results and they should be considered with caution when it comes to time-series data.

In addition, in Kelly and Pruitt (2015) the forecast of a single time series using many predictor variables is done with a new estimator called the three-pass regression filter (3PRF). According to the authors, the 3PRF demonstrates strong forecasting performance, and is often superior to alternatives, across a variety of simulation specifications and in empirical applications using macroeconomic and financial data. This new estimator, 3PRF, is nothing but a constrained least squares estimator and reduces to Partial Least Squares as a special case.

3 Data and methodology

We use data from the monthly macroeconomic database from the federal reserve bank of St. Louis (McCracken & Ng, 2016). This database consists of 127 monthly US indicators [1] from January 1959 to September 2022. We focus our analysis on pre-Covid years, so we drop the indicators after december 2019.

The dataset consist of series on industrial production and income, on labor market indicators, housing, consumption and inventories, money and credit, interest and exchange rates, prices, and the stock market.

Series are transformed to obtain stationarity following the recommendations from McCracken and Ng (2016). [2] All our transformations are annual. However, consistently with De Mol et al. (2008), for real variables, such as employment, industrial production and sales, we take the (annual) growth rate, while, for series already expressed in rates, as unemployment rate and capacity, we take (annual) first differences.

In addition, an important feature related to our dataset should be highlighted: the variable "Reserves Of Depository Institutions" ("NONBORRES", see figure 11) displays in January 2010 outliers anomalous with respect to the scale of the whole series.

As De Mol et al. (2008), we focus our attention on the forecast of the consumer price index (CPI), henceforth π_t. To obtain stationarity, we take the difference of the growth rates.

[1] Of the 127 series we use only 121 because of missing values as we try to preserve the highest number of observations possible. The variables that are dropped are: New Orders for Consumer Good (missing values until 1992), New Orders for Non-defense Capital Goods (missing until February 1968), VIX (missing until July 1962), Trade Weighted U.S. Dollar Index (missing until December 1972), Consumer Sentiment Index (missing until December 1972), Reserves Of Depository Institutions.

[2] However, the variables "SRVPRD", "USFIRE", "USTPU", "USTRADE", "CONSPI", "HWI", "USCONS", "BUSINVx", "USWTRADE", "UEMP27OV", "PAYEMS" displays a p-value for the Dickey-Fuller test, under the null of non-stationarity, greater than 0.010.

The accuracy of predictions is, as De Mol et al. (2008), evaluated using the mean-square forecast error:

$$MSFE^h = \frac{1}{T_1 - T_0 - h + 1} \sum_{t=T_0}^{T_1-h} (\hat{\pi}_{t+h|t+h-1} - \pi_{t+h})^2$$

Our final sample has a monthly frequency and ranges from 1961:01 to 2019:12. The evaluation period is 1971:01 to 2019:12. $T_1 = 2019{:}12$ is the last available point in time, $T_0 = 1970{:}12$ and $h = 12$. In detail, in the next sections, for each method, we report MSFE relative to the MSFE of the random walk (i.e., we use the one-year-before observation as a prediction)

$$MSFE^h_{\text{RW}} = \frac{1}{T_1 - T_0 - h + 1} \sum_{t=T_0}^{T_1-h} (\pi_t - \pi_{t+h})^2 \implies MSFE^h_{\text{R}} = \frac{\sum_{t=T_0}^{T_1-h} (\hat{\pi}_{t+h|t+h-1} - \pi_{t+h})^2}{\sum_{t=T_0}^{T_1-h} (\pi_t - \pi_{t+h})^2}$$

and for three sub-periods: the first two 1971–1984 and 1985–2003 are exactly the same as De Mol et al. (2008) but, in addition, since we enlarge the evaluation period, we add a third subsample 2004-2019. Given the results of De Mol et al. (2008) and the fact that the predictability of key macroeconomic time series has dramatically decreased, we indeed expect that the relative performance of the four methods will decrease going from the first subsample to the third.

We use rolling estimates with a 10 years window: at each t, we use the previous 10 years to estimate the parameters of a linear model and we use it to obtain a prediction of the dependent variable in the first month after the rolling window

$$\hat{\pi}_{T+h+1|T+h} = \boldsymbol{x}'_{T+1}\boldsymbol{\beta}$$

where T is the last observation in time of the predictors for a given rolling window. It is worth highlighting the presence of π_t among the predictors $\boldsymbol{x}'_t$. Notice that in each rolling window we use 11 years of information, due to the 12 month skip in the forecast, e.g. the first rolling window uses the predictors going from 1961:01 to 1970:12 and the dependent variable going from 1962:01 to 1971:12 as training, and we get a prediction of π_{T+h+1} at 1972:01, using the values of the predictors at 1971:01.

Since in each rolling window we have 120 months of observations and 121 variables, OLS is not feasible. Even if it were feasible ($T > n$), since we have a lot of predictors and not so many points in time ($T \approx n$), the OLS will tend to overfit and the quality of the predictions would not be good. Therefore, we estimate $\boldsymbol{\beta}$ using 4 different methods: PCR, PLS, Ridge and LASSO.

Notice that we do not standardized the whole dataset during the data cleaning phase. Instead, we standardize the predictors inside each rolling window. This has two reasons:

- The first is mathematical: since we want to use standardized variables in the estimations, we have to standardized inside each rolling window.
- The second is about information: by standardizing the whole dataset at the beginning, we are using information that is not in the information set of each rolling window (i.e. the overall means and variances are known only if we assume that the information set contains the whole time period).

When it comes to standardizing the test predictors, $\boldsymbol{x}_{T+1}$, we cannot compute the mean and the variance of a single row of predictors, since the former is exactly the value and the variance is null. Therefore, in order to achieve this standardization, we use the means and the standard deviations of the predictors' training sample, inside each rolling window.

4 PCR

4.1 Theory

Principal components regression is an interesting and useful way of overcoming the infeasibility of the OLS when we are working in high dimensions. The principle is to create $M < T$ linear combinations of the predictors and regress the dependent variable on them. In PCA, we choose the linear combinations as the orthogonal directions along which the data varies the most. The sample variance of a given linear combination $\boldsymbol{z}_1$ of the predictors with normalized weights (i.e. $\boldsymbol{\phi}_1'\boldsymbol{\phi}_1 = 1$) is:

$$\frac{\boldsymbol{z}_1'\boldsymbol{z}_1}{n} = \frac{\boldsymbol{\phi}_1' X' X \boldsymbol{\phi}_1}{n}$$

The first principal component is defined as:

$$\boldsymbol{v}_1 = \underset{\boldsymbol{\psi}}{\operatorname{argmax}} \frac{\boldsymbol{\psi}' X' X \boldsymbol{\psi}}{n}, \quad \text{s.to } \boldsymbol{\psi}'\boldsymbol{\psi} = 1$$

It is straightforward that the solution of this maximization satisfies:

$$\frac{X'X}{n}\boldsymbol{v}_1 = \lambda \boldsymbol{v}_1$$

where λ is the Lagrange multiplier of the maximization problem. Therefore, $\boldsymbol{v}_1$ is the eigenvector corresponding to the highest eigenvalue of the covariance matrix of X (recall that X is standardized and has 0 mean). Moreover, the m-th PC is just the eigenvector corresponding to the m-th eigenvalue of the covariance matrix of X.

PCA can be very useful when there is a lot of correlation between the predictors, and that is one of the characteristic features of time series data (see figure 12). The intuition behind the statistical utility is that there are some factors that drive the comovement between the variables and that they can be reconstructed as linear combinations of the data.

Therefore, the final model is:

$$\boldsymbol{y} = \beta_0 + \sum_{j=1}^{M} \beta_j X \boldsymbol{v}_j + \boldsymbol{e}$$

and we can estimates $\boldsymbol{\beta}$ by OLS (assuming we choose $M < T$) as:

$$\hat{\boldsymbol{\beta}}_M^{\text{PCR}} = (Z'Z)^{-1} Z'\boldsymbol{y}$$

where the columns of Z are the first M principal components and $\boldsymbol{\iota}$ (which is a vector of ones, used to estimate β_0).

4.2 Empirics

As explained in the preivous section, we run PCR inside each rolling window (after having standardized the training predictors) and we make a prediction of the transformed CPI for the first month outside the window.

We assess the forecasts' accuracy for 7 different numbers of principal components. Figure 1 clearly shows that using the first 75 PCs seems to minimize the out-of-sample prediction error. Table 1 shows the MSFE for all the numbers of principal components and for the different subperiods. We can notice how our results confirm what De Mol et al. (2008) showed about the predictability of macro variables: the MSFE relative to the random walk's one is significantly lower in the first subperiod than in the last two. However, it seems that the predictability of the series is higher in the last period than in the second. Yet, by comparing the forecasted series and the true one, we can notice from figure 2 that

	Number of components						
	1	5	10	25	50	75	100
1971-1984	0.353	0.222	0.158	0.078	0.054	0.044	0.071
1985-2003	0.434	0.443	0.342	0.163	0.124	0.096	0.142
2004-2019	0.339	0.253	0.246	0.133	0.107	0.089	0.234
1971-2019	0.375	0.283	0.230	0.117	0.089	0.073	0.153

Table 1: PCR MSFE relative to random walk for different subperiods

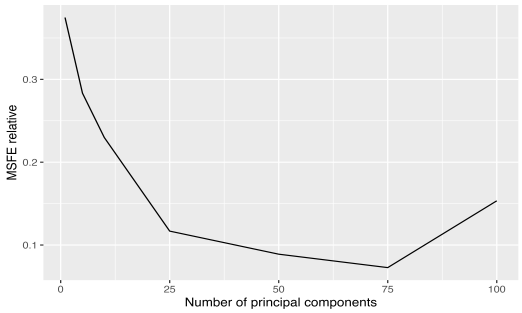

Figure 1: PCR MSFE relative to random walk

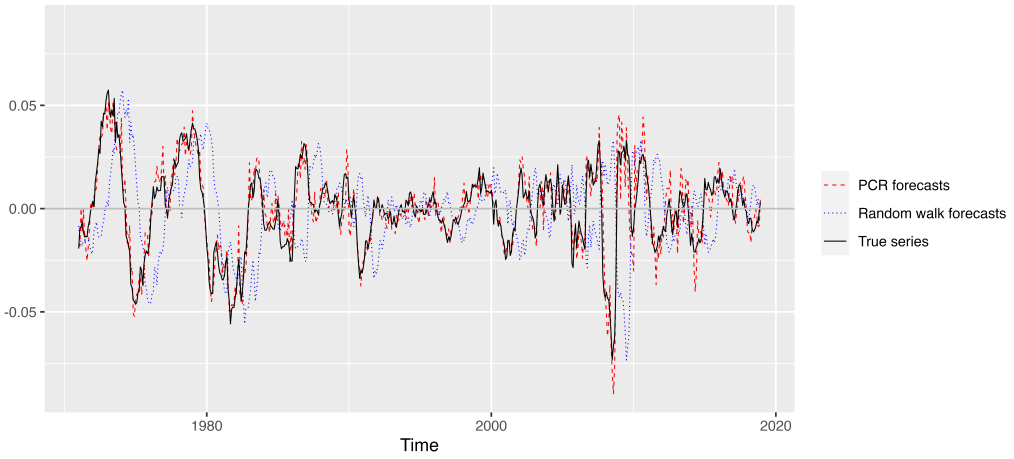

Figure 2: PCR forecasts with 75 principal components, random walk forecasts and true series

it might be the case that even the random walk has decreased its predictability power (mainly due to the high variability of the financial crisis), hence the relative MSFE increased even if the variables are still as predictable as the second period.

Figure 3 shows the scree plot (for the whole sample). Notice that there are 3 big gaps at the beginning, yet there is a small, but noticeable, one between the seventh and the eighth eigenvalues.

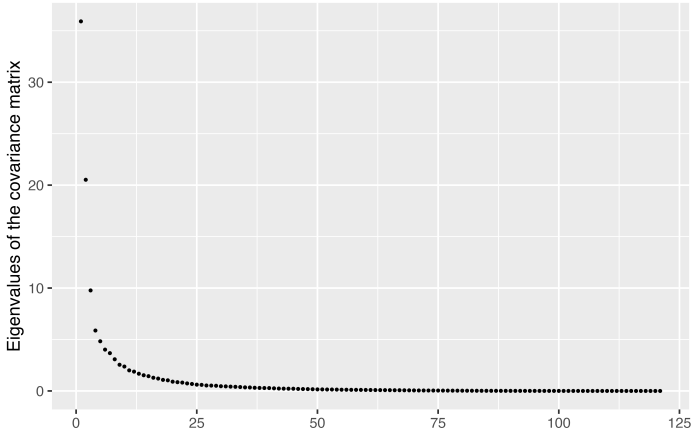

Figure 3: Scree plot for the whole dataset

From this, we conclude we do not need 75 components (as the MSFE suggests), yet the number should be between 1 and 10. However, we will report the correlation between the forecasts that minimized the relative MSFE for the PCR and the forecasts obtained with the other methods, to remain consistent with De Mol et al. (2008).

5 PLS

5.1 Theory

A drawback of PCR is that it does not ensure that the direction that best explains the variation of the regressors is also the one that best explains the direction of $\boldsymbol{y}$. Partial least squares is a supervised learning alternative to PCR.

PLS consists of assigning weights, $\boldsymbol{\alpha}_1' = (\alpha_{11}, ..., \alpha_{n1})$, to each regressor depending on how much variation of $\boldsymbol{y}$ they explain. Each α_{j1} is the OLS coefficient of regressing $\boldsymbol{y}$ on $\boldsymbol{x}_j$:

$$\alpha_{j1} = (\boldsymbol{x}_j'\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j'\boldsymbol{y} \tag{1}$$

Then, the first linear combination (the first direction) of the predictors is:

$$\boldsymbol{z}_1 = X\boldsymbol{\alpha}_1 = \sum_{j=1}^{n} \boldsymbol{x}_j \alpha_{j1} \tag{2}$$

The weights for the second linear combination are obtained by regressing $\boldsymbol{y}$ onto the residuals from regressing $\boldsymbol{x}_j$ onto $\boldsymbol{z}_1$, and so on. Like PCR, PLS identifies a set of M linear combinations of the regressors, as explained above, and then fits $\boldsymbol{y}$ via least squares using the M new regressors.

$$\hat{\boldsymbol{y}} = Z(Z'Z)^{-1}Z'\boldsymbol{y} \tag{3}$$

5.2 Empirics

We now perform PLS to again predict CPI. We do so with the same number of components as we did in PCR. In table 2 we present the mean squared forecast errors relative to the random walk for different sub-samples and number of components, and the correlation of the PLS forecast on the full sample with those of PCR with 75 components.

PLS consistently outperforms the random walk across different number of components and different subsamples. This advantage with respect to the random walk is larger as the number of components grows towards 25, but then becomes smaller as the number of components grow beyond 25. This U-shaped MSFE can be seen clearly in figure 4.

Across sub-samples, PLS performes best in the first period (1971-1984), and worst in the second. For all periods, it follows the same pattern of MSFE across the different number of components, and minimizes the MSFE at 25 conponenets.

Regarding the correlation with the PCR forecasts (with 75 principal components), the maximum is achieved with when using 25 components, and the better that PLS performs, the higher this correlation is. PLS outperforms PCR for lower number of components but PCR performs better when the number of components is higher.

	Number of components						
	1	5	10	25	50	75	100
1971-1984	0.218	0.077	0.054	0.044	0.07	0.15	0.214
1985-2003	0.402	0.214	0.127	0.085	0.125	0.197	0.465
2004-2019	0.269	0.145	0.106	0.075	0.145	0.206	0.417
1971-2019	0.286	0.131	0.089	0.065	0.112	0.183	0.342
Correlation with PCR (PC=75)	0.622	0.85	0.932	0.993	0.939	0.865	0.763

Table 2: PLS MSFE relative to random walk for different subperiods

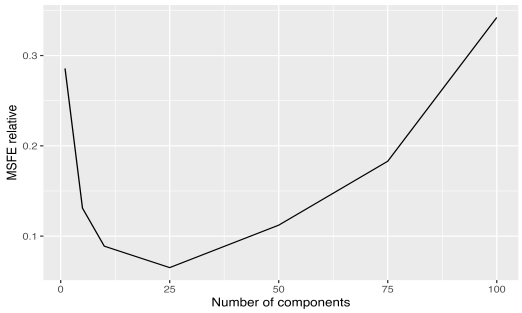

Figure 4: PLS MSFE relative to random walk

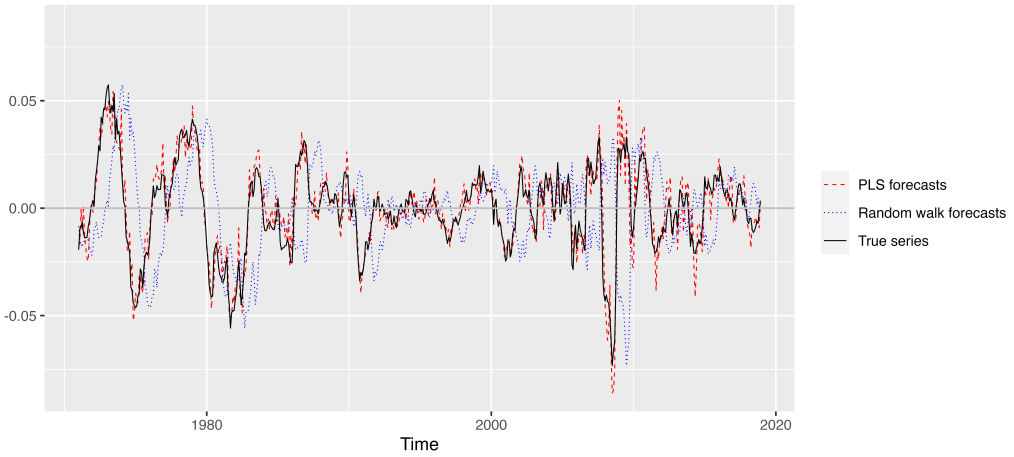

Figure 5: PLS forecasts with 25 components, random walk forecasts and true series

6 Ridge

6.1 Theory

Assuming a linear model $y_t = \beta_0 + \sum_{j=1}^{n} x_{tj}\beta_j + u_i$, the Ridge estimator is defined as:

$$\hat{\boldsymbol{\beta}}_{\lambda}^{R} = \underset{\boldsymbol{b}}{\operatorname{argmin}} \sum_{t=1}^{T} \left(y_t - b_0 - \sum_{j=1}^{n} b_j x_{jt} \right)^2 + \lambda \sum_{j=1}^{n} b_j^2 \tag{4}$$

Indeed, the Ridge estimator is the minimizer of the Residual Sum of Squares plus a penalty that depends on the tuning parameter λ. According to the level of penalisation λ, Ridge regression pushes the coefficients close to 0 (but not exactly 0, as Lasso does). Solving the problem written above, it can be easily shown that the Ridge estimator has a closed form solution:

$$\hat{\boldsymbol{\beta}}_{\lambda}^{R} = (X'X + \lambda I)^{-1} X'\boldsymbol{y} \tag{5}$$

which consists of adding a perturbation defined by λI to $X'X$ that allows $X'X + \lambda I$ to be always positive definite and, consequently, invertible [3]. Notice that X contains also a constant term (which is not standardized), which is needed to estimate the intercept. Indeed, $\hat{\boldsymbol{\beta}}_{\lambda}^{R}$ is a family of solution indexed by λ: as λ changes, the solution of the Ridge minimization problem changes. By increasing λ, the weight of the penalty increases and the flexibility of Ridge decreases: all the coefficients are,

[3] Indeed, when $n > T$, $X'X$ is not invertible and OLS is unfeasible. It can be easily show that $X'X + \lambda I$ is always positive definite and hence invertible for every possible value of λ greater than 0.

in fact, pushed towards 0 and the bias increases. On the other hand, because of the bias-variance trade-off, the variance decreases.

6.2 Empirics

For the Ridge regression, we perform the same exercise done for PCR. We start from our dataset after all the transformations described in section 3, including the standardization within rolling windows. As a matter of fact, since Ridge has the sum of squared coefficients in the penalty, its predictions are not scale invariant and require standardized predictors.

We analyse the performance in forecasting (with respect to the random walk) the CPI over a grid of λ. Table 3 shows the MSFE relative to the random walk over 7 values of λ and the correlation with the PCR forecasts. Looking at the overall evaluation period 1971-2019, as De Mol et al. (2008) we find that for the smallest value of λ Ridge performs worse. In our case it performs even worse than the random walk. We recall that for $\lambda = 0$ Ridge is exactly OLS, if it were feasible (i.e. $T > n$), since the minimization problem simply becomes minimizing the Residual Sum of Squares. With such small $lambda = 10^{-6}$, Ridge tends to overfit, which explains the MSFE obtained.

As we can also see from figure 6, overall the lowest relative MSFE is for intermediate values of λ: as λ increases the in-sample residual variance increases and, consequently, in an opposite reasoning with respect to the case of λ close to 0, the MSFE increases. Indeed, for the λ that minimizes the MSFE over the whole evaluation sample, we can plot the Ridge forecasts for the variable of interest together with its actual series and the Random Walk forecasts, as done in the previous sections, appreciating how well Ridge performs overall. Table 3 above, however, shows how the ability to predict has declined (not dramatically in our case) over time (for any value of λ), as also De Mol et al., 2008 finds. Moreover, this appears to be an inherent feature also of the most recent period (2004-2019), even if showing MSFE lower than the intermediate subperiod, probably due to the lower forecasting power of the Random Walk.

The last line of table 3 shows the correlation between Ridge forecasts and Principal Component forecasts. The two forecasts are highly correlated for all values of λ. The lowest correlation is for the value of λ (10^{-6}) for which Ridge delivers bad forecasts while the correlation is maximal for parameters giving the best forecasts. This suggests that there is a common explanation for the good performance of the two methods. As a matter of fact, since the covariance of our data is characterized by few dominant eigenvalues (because we have high collinearity among predictors; recall figure 3), PC and Ridge forecasts behaves really similarly. Indeed, while PC keep only the largest ones, Ridge, acting as a penalized PCR, keeps all of them but with decreasing weights.

	Lambda						
	10^{-6}	10^{-4}	10^{-2}	1	10^{2}	10^{4}	10^{6}
1971-1984	0.786	0.229	0.086	0.055	0.114	0.441	0.668
1985-2003	1.588	0.441	0.16	0.106	0.225	0.37	0.393
2004-2019	2.463	0.44	0.129	0.066	0.118	0.288	0.326
1971-2019	1.653	0.356	0.12	0.071	0.142	0.366	0.473
Correlation with PCR (PC=75)	0.434	0.755	0.929	0.966	0.822	0.613	0.592

Table 3: Ridge MSFE relative to random walk for different subperiods

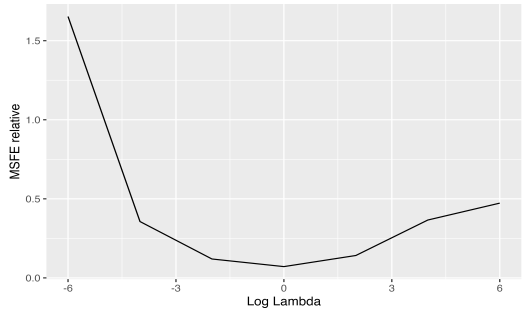

Figure 6: Ridge MSFE relative to random walk

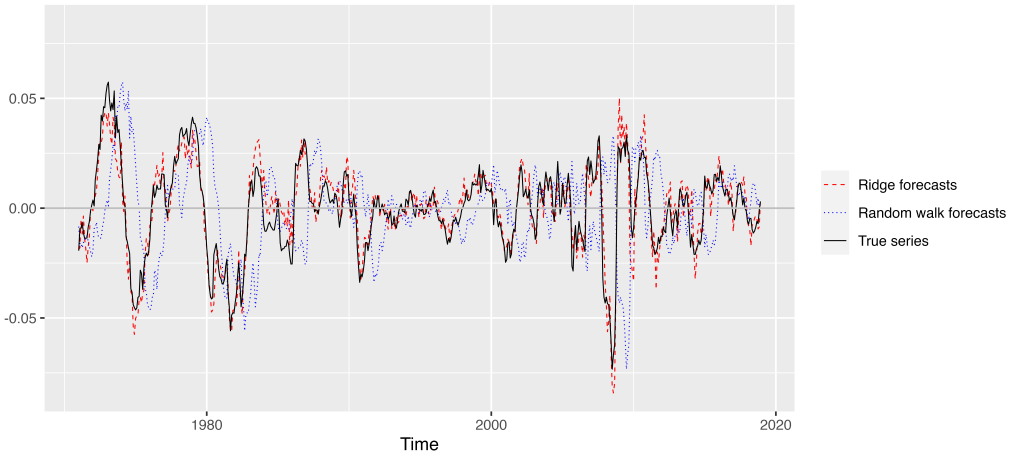

Figure 7: Ridge forecasts with $\lambda = 1$, random walk forecasts and true series

7 LASSO

7.1 Theory

Ridge has a disadvantage: it will include all the predictors in the model. This is due to the fact that the penalty in equation (4) will shrink all the coefficients towards zero but it will not reduce any of them exactly to zero. Therefore, Ridge does not perform variable selection [4]. While this may not be a problem when discussing prediction accuracy, it could pose challenges in the context of model interpretation.

The LASSO (Least Absolute Shrinkage and Selection Operator) coefficients, $\hat{\boldsymbol{\beta}}_\lambda^L$, are defined as:

$$\hat{\boldsymbol{\beta}}_\lambda^L = \underset{\boldsymbol{b}}{\operatorname{argmin}} \sum_{t=1}^{T} \left(y_t - b_0 - \sum_{j=1}^{n} b_j x_{jt} \right)^2 + \lambda \sum_{j=1}^{n} |b_j| \tag{6}$$

where $\lambda > 0$ is again the tuning parameter needed for regularization. As we can see, LASSO uses an l_1 penalty, in contrast to Ridge, which uses an l_2 penalty. Since l_1 penalty is not differentiable in $\boldsymbol{b} = \mathbf{0}$, this implies that a close form solution as we found for Ridge does not exists. This type of penalty has the effect of driving some coefficient estimates to become exactly zero when the tuning parameter is large. This allows us to perform variable selection and results in a model that is easier to interpret. This type of models are called sparse models. Therefore, as λ increases, the flexibility of LASSO decreases. This leads to an increase in bias but a decrease in variance. Conversely, as $\lambda \to 0$, bias decreases and variance increases.

However, in the time series setting, since variables are highly correlated, we expect variable selection to be unstable. In this context, variable selection is unlikely to yield results with clearer economic interpretations compared to principal components regression or Ridge regression, as all coefficients have an equal probability of being selected.

7.2 Empirics

We proceed to do the same as we did for the other methods, starting from the transformed dataset, using the function *glmnet* on the standardize (within rolling windows) predictors. The results that we obtain are summarized in table 4, which presents the mean squared forecast error relative to those of the random walk for 7 different levels of penalty and different sub-samples and the correlation

[4] This will happen if and only if $\lambda \to \infty$

	Lambda						
	10^{-3}	10^{-2}	10^{-1}	1	10	10^{2}	10^{3}
1971-1984	0.702	0.167	0.069	0.04	0.043	0.048	0.049
1985-2003	0.413	0.323	0.153	0.084	0.086	0.094	0.096
2004-2019	0.339	0.204	0.091	0.068	0.103	0.121	0.123
1971-2019	0.393	0.16	0.078	0.061	0.083	0.09	0.091
Correlation with PCR (PC=75)	0.437	0.803	0.941	0.985	0.973	0.965	0.964

Table 4: LASSO MSFE relative to random walk for different subperiods

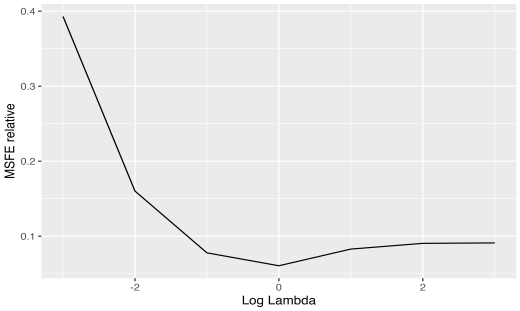

Figure 8: LASSO MSFE relative to random walk

with PCR. Similar to De Mol et al. (2008), we begin by analyzing the overall evaluation period. We discover that LASSO consistently outperforms the random walk across various penalty values. This advantage becomes more significant as the penalty value increases and reaches its peak when λ equals 1; after this point, the relative MSFE starts to increase again. As already shown in the other methods, the predictability seems to have dropped from 1985 onwards, thus, the same comment of the previous sections applies.

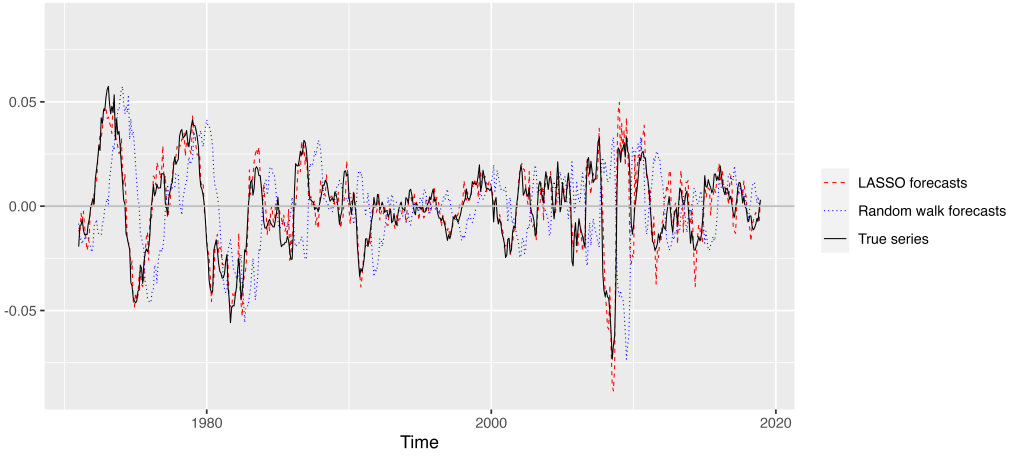

Figure 9: LASSO forecasts with $\lambda = 1$, random walk forecasts and true series

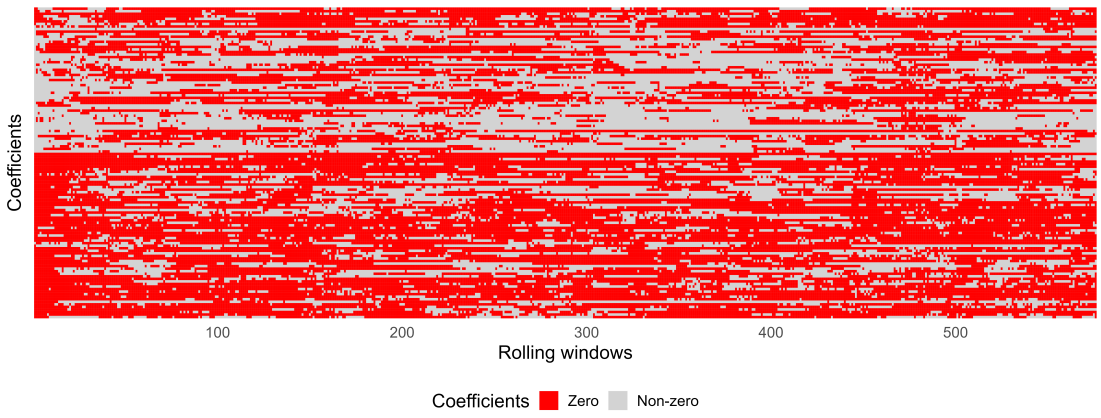

Figure 10: Coefficients shrunk by LASSO with $\lambda = 1$ in each rolling window

The last row of table 4 represents the correlation with PCR, which increases as lambda becomes larger until a certain point where it reverses. As we can observe, LASSO and PCR exhibit a high level of correlation, primarily attributed to the high collinearity within our data. In situations of strong data correlation, only a small subset of variables may capture the data's covariation, similar to what principal components do.

However, as we said before, in this setting, this selection is not stable and is sensitive to minor data perturbations, meaning that all the coefficients have the same probability of being shrunk. We can see this for our case in figure 10. This figure shows whether LASSO coefficients (in the y-axis) are shrunk to zero (in red) or not (in light gray) for each rolling window (x-axis). The coefficients that are being shrunk to zero are not consistent across rolling windows.

8 Results and conclusions

This paper has analyzed the forecasting performance (relative to Random Walk) of different Linear Supervised Learning Methods, using as benchmark Principal Component Regression. Both PLS and shrinkage methods (Ridge and Lasso) offer a valid alternative to PCR: indeed, as in De Mol, the forecasts provided are highly correlated to those of PCR.

Overall, all methods outperform the random walk and show U-shaped MSFEs, as expected from the underlying bias-variance trade-off that characterizes different level of regularization. Indeed, the plots obtained using the optimal level of regularization for each method, shown in the previous sections, display how closely the forecasts are to the true series of the CPI. Nevertheless, the predictive power (considering again the optimal λ and number of components for each method) has declined over time: the MSFE (relative to Random Walk) is higher in the sub-periods 1985-2003 and 2004-2019 with respect to 1971-1984. It is important to highlight that MSFE has decreased between the second and the third subperiod, suggesting that the forecasting power of the predictors has risen again in the most recent years. However, as it is possible to see from the forecasts plots, the decrease in the relative MSFE is probably due to a worse forecasting power of the random walk (because of huge shocks within short time periods during the financial crisis), rather than a better performance of our methods.

Finally, an important warning regarding LASSO must be stated here again. Being a variable selection method, it is unlikely to provide results that are more interpretable than Principal Components, Ridge or Partial Least Squares regressions from the economic point of view, even if we obtain similar forecasting performance.

Division of work

- Introduction: All
- Literature review: All
- Data and methodology: Pablo Suarez-Sucunza and Iari Orlandi (Data cleaning), Samuele Borsini and Mario D'Agostino (Estimation strategy)
- PCR: Samuele Borsini
- LASSO: Iari Orlandi
- Ridge: Mario D'Agostino
- PLS: Pablo Suarez-Sucunza
- Conclusion: All

Bibliography

De Mol, C., Giannone, D., & Reichlin, L. (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, *146*(2), 318–328.

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

Kelly, B., & Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, *186*(2), 294–316.

McCracken, M. W., & Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589.

Appendix A: *NONBORRES* time series

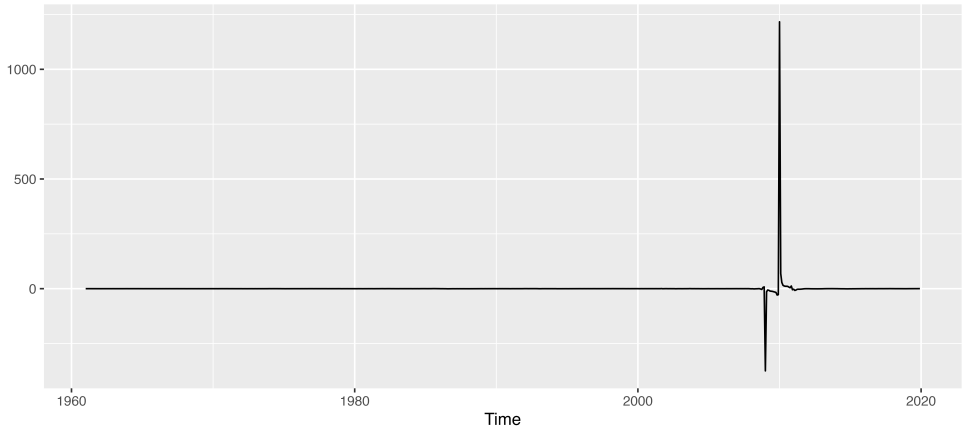

Figure 11: Variable *NONBORRES* transformed

Appendix B: Correlation among regressors

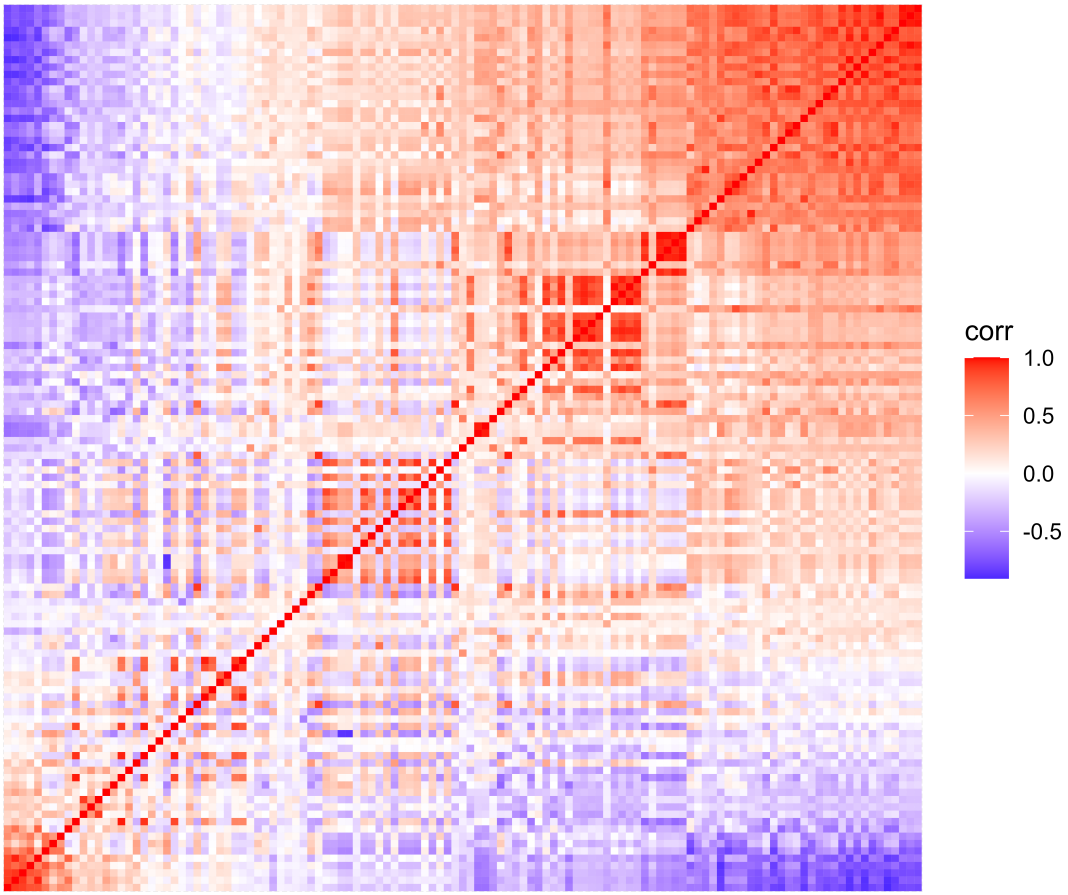

Figure 12: Correlation heatmap of the regressors