,

Forecasting with many predictors: a comparison among Ridge, Lasso, PCA, and 3PRF

Mino Brescia 0001088560

James Daniel Foltz 0001076043

Francesco Di Marzio 0001084497

Farzad Yaghoobi 0001085576

Editor:

Abstract

This project explores and compares the predictive power of four statistical learning methodologies: Bayesian regression with normal and double-exponential priors, the Three Pass Regression Filter (3PRF), and Principal Component Analysis (PCA). These methods are evaluated in the context of forecasting, particularly when dealing with large panels where the number of time series is not small relative to the number of observations. In agreement with previous analysis of the issue, we did an out-of-sample evaluation based on the McCracken dataset (McCracken and Ng, 2015). The project also delves into the work of Stock and Watson, which considers the efficiency of forecasts when many predictors and time series observations are involved (Stock and Watson, 2002). We see that these predictors can be condensed into a small number of indices using principal components.

Keywords: Factor Model, Forecasting, Principal Components, Ridge, Lasso, 3PRF

1. Introduction

Economic research frequently relies on studying extensive collections of time series data. Recent advances in this field have highlighted the significant benefits of using large datasets for extracting meaningful signals and making accurate forecasts. To tackle the complexities associated with handling these high-dimensional datasets, new techniques have been introduced (Forni et al., 2005; Giannone et al., 2004; Stock and Watson, 2002a,b). These are statistical learning methods which have been investigated by the scholars. The literature on principal components and classical factor models, for instance, is large and well known (Lawley and Maxwell, 1962).

In the realm of forecasting, Bayesian model averaging has garnered significant attention (Koop and Potter, 2003; Stock and Watson, 2006, 2005a; Wright, 2003). Notable works include Stock and Watson's (2005a) investigation of normal Bayes estimators for orthonormal regressors and Giacomini and White's (2006) empirical comparison of Bayesian regression to principal component regression (PCR).

The De Mol paper (Mol et al., 2008), with its insights into Bayesian regression under Gaussian and double-exponential priors, serves as a basis for our project. It delves into the performance of Bayesian regression methods and their potential to overcome the curse of dimensionality by leveraging priors. It also draws connections between Bayesian regression and the classical literature on forecasting with large panels based on principal components. This work sets the stage for our project to evaluate and compare the forecasting performance of Ridge, Lasso, and PCA.

In parallel, the Three Pass Regression Filter (3PRF) paper by (Kellya and Pruitt, 2015) introduced an innovative estimator designed for single time series forecasting using a multitude of predictor variables. 3PRF calculations are straightforward, represented as sets of ordinary least squares regressions, which offer the potential for consistent forecasts as both the dimensions of the time series and cross-sectional data expand. This novel approach is highly promising, and its implementation is a vital part of our comparative study.

Our research objectives align with De Mol's exploration of Bayesian regression but extend the analysis to encompass this additional method. Our aim is to evaluate the forecasting capabilities of these techniques when applied to macroeconomic panels. Furthermore, our project will explore the applicability of these methods to real-world forecasting scenarios and evaluate how do they perform.

Our endeavor is structured as follows: Section 2 introduces the challenges associated with forecasting using extensive cross sections (the "curse of dimensionality" problem), drawing insights from De Mol's and 3PRF's contributions. Section 3 will showcase our empirical results, comparing the forecasting performances of Ridge, Lasso, PCA, and 3PRF on the macroeconomic dataset. Finally, in Section 4, we will conclude our study.

2. Theoretical considerations: four different methods

The "Curse of Dimensionality" is a concept in statistical learning and data analysis that refers to the various challenges and problems that arise as the dimensionality (the number of features or variables) of a dataset increases. When the size of the information set (n) is large, the OLS projection of the dependent variable on all the available time series involve

the estimation of a large number of parameters. This implies a loss of degrees of freedom and a poor forecast, along with being computationally intensive. Another problem tied to the high dimensionality of the predictive information is overfitting, which means that models may capture noise or spurious patterns in the data rather than true underlying relationships and sacrifice interpretability to gain more flexibility. In many high-dimensional datasets, only a small fraction of features may be relevant to the problem, requiring effective feature selection techniques to identify and focus on the relevant ones. Moreover, if the number of regressors is larger than the sample size, n(p + 1) > T, OLS is not feasible. Some statistical learning methods have been suggested and used in order to deal with these problems. Here we apply four of these methods which are suggested in the literature. In what follows we explain the theory of each method briefly.

2.1 Lasso

The Lasso (Least Absolute Shrinkage and Selection Operator) corresponds to a Bayesian regression in which the priors follow a double-exponential zero mean i.i.d. distribution. According to the oracle property, the lasso correctly identifies the true non-zero coefficients as the sample size grows to infinity and it is consistent and computationally efficient. These properties make it useful for causal inference. Although it may not always achieve the oracle property, the Lasso is still useful as a feature selection and dimensionality reduction algorithm. The Lasso is a method used to select the most relevant variables to explaining the target when the available number of regressors is large compared to the number of observations. The lasso parameters are estimated by minimizing the sum of squared residuals subject to the constraint that the sum of the absolute value of the coefficients be less than a prespecified threshold. The lower the threshold, the farther the estimates will be from the OLS ones. The problem can be stated as

$$\hat{\beta_s^L} = \arg\min_{\mathbf{b}} RSS(\mathbf{b}) \;\; s.t. \;\; \|\mathbf{b}\|_1 \leqslant s \tag{1}$$

The problem can be represented for two regressors by plotting them in a bidimensional space. The constraints will be represented by a square whose sides are defined by $\|\beta_1\| + \|\beta_2\| = s$. Differently from Ridge, OLS, and PCR, Lasso will shrink some coefficients to zero (it favors sparse regression coefficients). Thus, its results are highly interpretable, although its performance tends to be poor while forecasting. The problem can be equivalently stated as

$$\hat{\beta^L} = \arg\min_{b} \|y - Xb\|^2 + \lambda \sum_{i=1}^{n} \|b_i\| \tag{2}$$

Where $\lambda \geqslant 0$ is a tuning parameter needed for regularization. When $\lambda = 0$ the coefficients coincide with the OLS estimates, while as $\lambda \to \infty$ all coefficients are shrunk to zero. Since OLS is unbiased, the bias of the Lasso forecast increases with λ, although its variance decreases. In our empirical exercise we try to determine the optimal Lambda by minimizing the estimated test MSE. Lasso has no closed form solution since its objective function is non-differentiable. Predictors need to be standardized and a solution is achieved through numerical techniques.

2.2 Ridge

Differently from the Lasso, the Ridge algorithm does not shrink the coefficients down to zero. To be more precise, It shrinks but retains all predictors (unless λ, the parameter in equation (4), is equals inf). The problem that Ridge solves is as the following:

$$\hat{\beta_s^R} = \arg\min_{\mathbf{b}} RSS(\mathbf{b}) \;\; s.t. \;\; \|\mathbf{b}\|_2^2 \leqslant s \tag{3}$$

Which is equivalent to:

$$\hat{\beta^R} = \arg\min_{b} \|y - Xb\|^2 + \lambda \sum_{i=1}^{n} b_i^2 \tag{4}$$

One of the advantages of Ridge over Lasso is that Ridge has a closed form mathematical solution, while Lasso does not. This is due to Ridge's objective function being differentiable. Ridge also uses a tuning parameter, λ as a penalty for regularization. It attempts to minimize the RSS while also shrinking coefficients. We can use Ridge when we want to keep all the variables in the model; however, it is more difficult to interpret the Ridge output compared to the output of Lasso.

2.3 Principal Components

PCA attempts to reduce the dimensionality in the data by selecting the directions in the space of the regressors along which the variance in the data is maximised. It is an unsupervised learning procedure since the selection procedure is not dependent on the target variable Y but only on the predictors X. The forecast is computed as a projection on the first few principal components. The components are each chosen so that they maximize the variance in the data subject to the constraint that they be orthonormal to each other. The spectral decomposition of the sample skedasiticty matrix of the predictors $S_x = \frac{1}{T-h-p} X'X$, is $S_x V = DV$, where $D = diag(d_1, ..., d_{n(p+1)})$ is the diagonal matrix of the eigenvalues of S_x in decreasing order of magnitude while $V = (v_1, ..., v_{n(p+1)})$ is the square matrix whose columns are the corresponding eigenvectors. The normalized principal components are defined as $\hat{f}_{it} = \frac{1}{\sqrt{d_i}} v_i' X_t$ for i = 1, ..., M, where M is the number of non-zero eigenvalues. When the interactions among variables in the dataset primarily result from a handful of shared underlying factors, and there is minimal cross-correlation among the unique components of the individual variables, it becomes possible to condense the wealth of predictors into a small number of aggregate variables. The unexplained portion, which is not accounted for by these common factors, can then be forecasted using traditional single-variable or low-dimensional forecasting techniques, often by projecting onto the dependent variable or a limited set of predictors. In such scenarios, a small number of principal components effectively represent the fundamental underlying factors. The principal component forecast is defined as: $F\hat{\theta}^{LS} = F(F'F)^{-1}F'y = \frac{FF'y}{n}$, where $F_t = (\hat{f}_{1t}, ..., \hat{f}_{Mt})$. Since $M << n(p+1)$, PCA achieves a sparse representation of the information set. The parsimonious approximation of the information set makes the OLS projection feasible, since it requires the estimation of a limited number of parameters.

2.4 Three Pass Regression Filter

While PCR condenses the cross section according to covariance within the predictors, some of the factors driving the panel of predictors may be irrelevant for the dynamics of the forecast target. The 3PRF does not use those factors but condenses the cross section according to the covariance with the forecast target. To achieve consistency the 3PRF need only estimate the relevant factors, which are always less than or equal to the total number of factors required by PCR. The 3-pass regression filter can be considered as an extension of the general dynamic factor model. We aim to forecast a target variable in the context in which we have available many predictors that we expect may contain information relative to the target variable. To reduce the dimension of the predictive information, we can assume the target variables are driven by a small number of factors that also drive the covariance among the predictors. To create predictions, the 3PRF relies on proxies. These proxies are variables that are influenced by the underlying factors, especially those that are relevant to the target variable. Proxies can be made available from economic theory or they can be selected through an automatic procedure from the target and predictor variables. The target variable is a linear combination of some of the hidden factors and unpredictable noise. The best possible forecast, therefore, would theoretically be achieved from a regression on the actual underlying factors that are relevant. However, because these factors are not directly observable, such forecast is unfeasible.
The 3PRF can be formulated as a sequence of OLS regressions. The first pass runs N separate time series regressions, one for each predictor. In the first pass regressions, the predictor is the dependent variable, the proxies are the regressors, and the estimated coefficients describe the sensitivity of the predictor to factors represented by the proxies.

$$x_{i,t} = \phi_{0,i} + \mathbf{z}'\boldsymbol{\phi}_i + \varepsilon_{i,t}$$

In the second step, we utilize the previously estimated coefficients from the initial pass in T separate cross-sectional regression analyses. In these secondary regressions, the independent variables remain the same as the original dependent variable, while the first-pass coefficients (represented as $\hat{\phi}_i$) act as predictors. The initial-stage coefficient estimates help us map the predictors across different cross sections to the latent factors. The subsequent cross-sectional regressions employ this relationship to obtain estimates of the factors at each time point.

$$x_{i,t} = \phi_{0,t} + \hat{\boldsymbol{\phi}}'\mathbf{F}_t + \varepsilon_{i,t}$$

These estimated second-pass predictive factors (denoted as $\hat{F}_t$) are then used in the third step. In this phase, we perform a single time series prediction using a regression model where the target variable y_{t+1} is forecasted based on the second-pass estimated predictive factors $\hat{F}_t$.

$$y_{t+1} = \beta_0 + \hat{\mathbf{F}}_t{}'\boldsymbol{\beta} + \eta_{t+1}$$

The fitted value $\hat{\beta}_0 + \hat{F}'\beta$ at the third stage represents the 3PRF forecast at time t. Given that the relevant factor space is spanned by $\hat{F}_t$, the third-stage regression yields reliable and consistent forecasts.

The target variable satisfies the requirements for the automatic proxy selection procedure in the case where there is only one relevant factor since it has zero loading on irrelevant factors, has linearly independent loadings on the relevant factors, and is in number equal to the relevant factors. If the number of relevant factors is greater than one, the target-proxy 3PRF falls short of asymptotically achieving the infeasible best.

In such a scenario, we can enhance the target-proxy 3PRF by introducing additional proxies that exclusively rely on relevant factors. We can generate a second proxy by observing that the residuals obtained from the target-proxy 3PRF forecasts also exhibit the previously mentioned properties. This is because they have non-zero loading on relevant factors (as a consequence of the inadequacy of the target-only proxy), zero loading on irrelevant factors (by definition), and remain linearly independent from the first proxy.

Subsequently, the proxy construction process unfolds iteratively, using the residual from the target-proxy 3PRF as the second proxy, the residual from this two-proxy 3PRF as the third proxy, and so on. When this iterative algorithm is employed to create M predictive factors, we refer to the forecaster as the M-automatic-proxy 3PRF.

3. Empirical analysis

3.1 Data

We use current monthly data from the McCracken dataset (McC) (McCracken and Ng, 2015). The dataset consists of a panel of macroeconomic variables whose monthly observations range from 01/03/1960 to 01/03/2023. The handling of missing data is as follows. In certain instances, specific variables within our dataset lacked data observations for the beginning period spanning from 1959 to 1960. In these cases, the entire monthly data for these observations was omitted, given the absence of information within those time frames. Additionally, some variables exhibited missing values for the most recent months within our dataset, so we opted to exclude the last 5 rows (from April to August 2023). In order to make our time-series stationary some transformations are made to the original dataset. The transformations are described in detail in the Appendix of (McCracken and Ng, 2015). According to them, the following data transformation has been made based on the nature of each time series x: (1) no transformation; (2) Δx_t; (3) $\Delta^2 x_t$; (4) $log(x_t)$; (5) $\Delta log(x_t)$; (6) $\Delta^2 log(x_t)$. (7) $\Delta(x_t/x_{t-1} - 1.0)$.

In general, for real variables, such as employment, industrial production, sales, we take the monthly growth rate. We take first differences for series already expressed in rates: unemployment rate, capacity utilization, interest rate and some surveys. Prices and wages are transformed to first differences of annual inflation.

Following previous studies, we take 1960:1 as the start of the sample. After losing two observations to data transformations (for some variables first differences are taken twice), the panel we use for analysis is for the sample 1960:3 to 2023:03 with 757 observations.

3.2 Econometric strategy

Lasso, Ridge and PCA models are used as described in the paper by Mol et al. (2008). For Lasso and Ridge, we produce an out-of-sample forecast for each month in the evaluation period by training the model on the ten previous years of data (for instance, y_t until y_{t+120}),

and we use the regression coefficients to estimate the target variable one year ahead (in this case y_{t+132}). This method generates one squared forecasting error for each window (computed as the square of the difference between the forecast and the actual observation). We then compute the average over the entire evaluation period to obtain the MSFE (Mean Squared Forecast Error) given a specific value of the regularization parameter (λ for Ridge and Lasso, M for PCA). The evaluation period ranges from 01/03/1971 to 01/03/2022. We choose Consumer Price Index (CPI) and Industrial Production (IP) as the target variables for our analysis. More specifically, the variables are transformed as follow: $\pi_t = 100\frac{logCPI_t}{logCPI_{t-12}}$ (annual CPI inflation), $ip_t = 100logIP_t$ (rescaled log of IP). We compute the MSFE using the following formulas:

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

$$MSFE_{ip}^{H} = \frac{1}{T_1 - T_0 - h + 1} \sum_{T=T_0}^{T_1-h} (\hat{ip}_{T+h|T} - pi_{T+h})^2$$

It is worth noting that we get an estimate of the MSFE for a specific value of the tuning parameter, and we then choose the optimal parameter to get the best forecast. Following De Mol's approach, we report the ratio between the MSFE estimated in this way and the one estimated with a naive random walk with drift. We fit a random walk to the target variables in each window by assuming that errors follow a normal distribution, with variance equal to the variance of the series in the training sample. The evaluation period lasts 143 months less than the available sample since we train the model on the first 10 years of data and the forecast is done one year ahead. We apply all the methods to stationary and standardized data.

The procedure for the PCA follows the one for Lasso and Ridge. More specifically, within each window, we estimate the matrix of principal component scores (Z) by multiplying the regressors (X) in the training sample by the matrix of the eigenvectors associated with the M largest eigenvalues. We then compute the principal component scores in the test sample by multiplying the regressors (out-of-sample) by the matrix of eigenvectors used to train the model.

Eventually, our goal is to apply the 3PRF mechanism in the rolling window setting described thus far, so as to evaluate how this new method perform relative to more standard techniques. This empirical exercise is not performed in the paper by Kellya and Pruitt (2015). We forecast using time windows starting at 5 arbitrary dates (1, 50, 100, 150, 200) to make the procedure less computationally intensive. We then iterate over chosen values of the number of factors M. In this case the test sample used to compute the MSFE corresponds to the 10 years of observations following the training sample. In addition, we fit a random walk with drift to get a 10-years ahead estimates. For both models, we compute the MSFE in the test sample and then compute the average of the MSFE for all the windows. We apply the procedure of the paper in the training sample to both get an estimate of the M proxies (matrix Z) and a vector of coefficients (β). In order to test the model we exploit Z matrix, the out-of-sample regressors and the vector of coefficients estimated in the training sample. This procedure leads to a prediction for the target variables for the entire 10 years timespan (and we then compute the MSFE based on the actual time series). The results are indicative of the prediction power of the 3PRF method in a rolling window setup.

3.3 Empirical Results

3.3.1 Lasso Results

Lasso regressions are estimated for different values of the regularization parameter λ. We report the ratio MSFE in the following table:

Table 1: Lasso forecasts

λ	0.005	0.02	0.029	0.04	0.1	0.2	0.3
Consumer price index							
MFSE 1960-2023	0.91	0.62	0.56	0.54	0.65	0.93	1.56
λ	4e-04	0.001	0.002	0.01	0.1	0.13	0.5
Industrial production							
MFSE 1960-2023	1.08	0.73	0.57	0.41	0.35	0.35	0.36

For the CPI the minimum ratio MSFE is 0.886 (λ=0.029) whereas for industrial production it is 0.35 (λ=0.01). As expected, the test MSFE features the usual U-shape. We further notice a substantial improvement over the random walk.

3.3.2 Ridge Results

We iterated the procedure to obtain the Ridge predictions for different values of the tuning parameter λ.

Table 2: Ridge forecasts

λ	0.04	0.1	0.26	0.4	1	3	5
Consumer price index							
MFSE 1960-2023	1.014	0.926	0.895	0.9	0.931	0.988	1.016
λ	0.002	1	8	30	50	70	100
Industrial production							
MFSE 1960-2023	1.017	0.327	0.288	0.308	0.322	0.332	0.34

The ratio MSFE is minimized for CPI around the value of lambda of 0.26, in which it equals 0.8945. For IP it is minimized at 8, in which it equals 0.288. Since the */lambda*'s are higher for IP rather than CPI, we conclude that the Ridge shrinks an higher number of parameters, and the dataset has less relevant variables to predict Industrial Production.

3.3.3 PCA Results

We report results for some numbers of principal components. The PCR performs better in forecasting CPI than a random walk when the number of principal components is between 58 and 76. In particular, the ratio is minimized at 0.945, for which the number of principal components is 68. The ratio is much lower when forecasting IP and is minimized at 5 components. The covariance among the predictors is summarized by a few principal components which also explain much of the variation in industrial production.

Table 3: PCR forecasts

PCs	10	30	50	60	68	80	90
Consumer price index							
MFSE 1960-2023	1.67	1.09	1.01	0.98	0.95	1.01	1.13
PCs	2	5	20	40	60	75	90
Industrial production							
MFSE 1960-2023	0.3	0.26	0.31	0.32	0.34	0.41	0.49

3.3.4 3PRF Results

Table 4: 3PRF forecasts

M	2	10	30	40	50	60	70
Consumer price index							
MFSE 1960-2023	0.395	0.5	0.46	0.46	0.48	0.49	0.51
M	5	10	15	20	25	30	35
Industrial production							
MFSE 1960-2023	0.28	0.27	0.29	0.30	0.31	0.33	0.33

The MSFE is minimized for CPI when the number of relevant factors is between 30 and 40, whereas it is minimized for IP when the number of factors is around 10. We again notice that the dimension of the predictive information is lower for IP than CPI.

3.3.5 Comparison

Our results suggest that Bayesian shrinkage method represents a better alternative to Principal component analysis for the variable Industrial Production (IP), whereas the same does not apply to Consumer Price Index (CPI). Regarding the 3PRF method, we notice a net improvement over the random walk, in particular for IP.

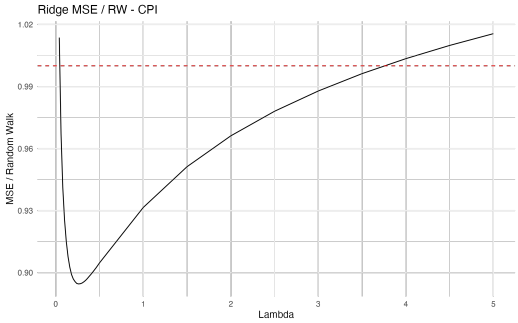
4. Conclusion

In our analysis, we conducted a comprehensive comparison of various forecasting methodologies, which included Ridge, Lasso, and PCA. Additionally, we showcased the practical application of the 3PRF method. This study provided valuable insights into the capabilities of each of these methods when applied to macroeconomic time series data.

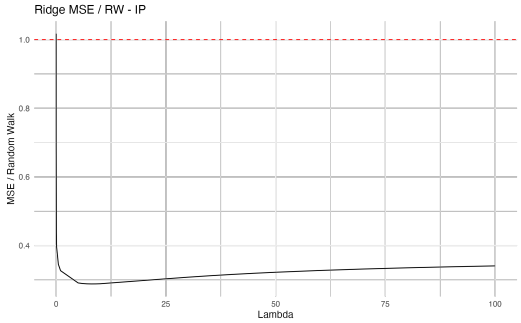
Our empirical findings clearly illustrate that Bayesian shrinkage methods, specifically Ridge and Lasso, serve as robust alternatives to Principal Component Analysis, especially in the context of forecasting the Consumer Price Index (CPI). On the other hand, when it comes to forecasting Industrial Production (IP), Principal Component Analysis demonstrated superior performance.

Furthermore, we demonstrated the potential utility of the 3PRF method in out-of-sample forecasting, underscoring its effectiveness. In conclusion, our study revealed that all four methods discussed in this project significantly outperform a basic random walk estimation. This analysis underscores the critical importance of selecting an appropriate theoretical framework for economic forecasting. By gaining an understanding of how different techniques perform when dealing with a multitude of predictors, we enhance our ability to obtain accurate and efficient economic forecasts and conduct insightful analyses.

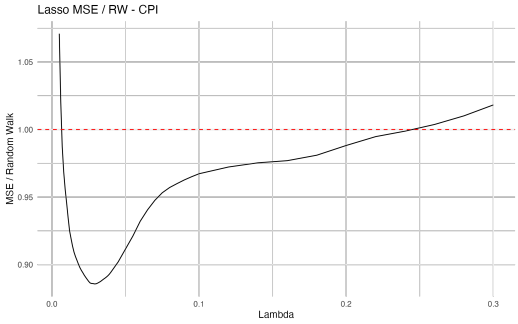
Appendix A. Figures



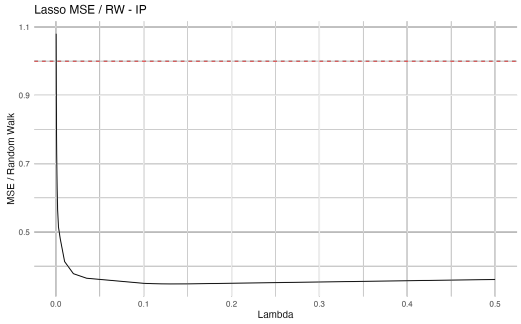
(a) Ridge MSFE - CPI



(b) Ridge MSFE - IP

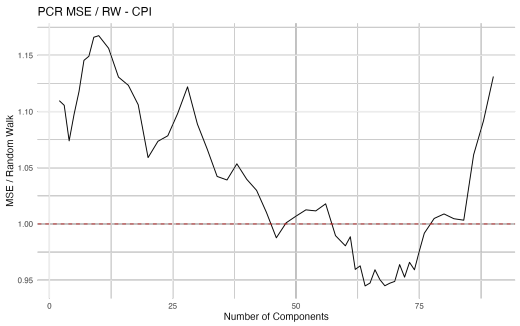


(c) Lasso MSFE - CPI

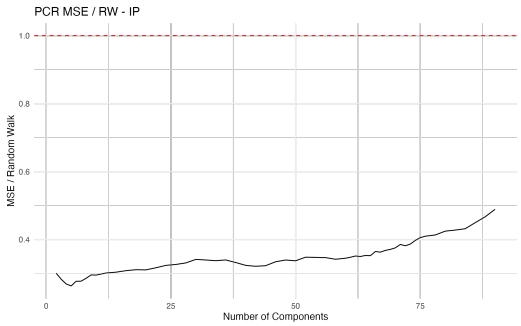


(d) Lasso MSFE - IP

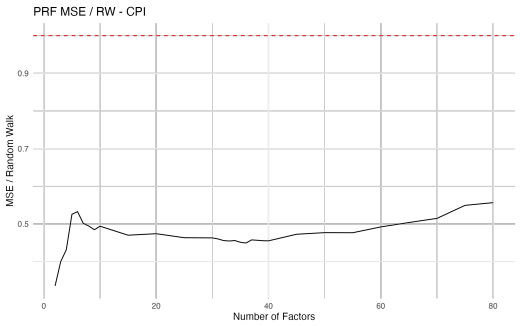
Figure 1: MSFE for the Ridge and Lasso and both series



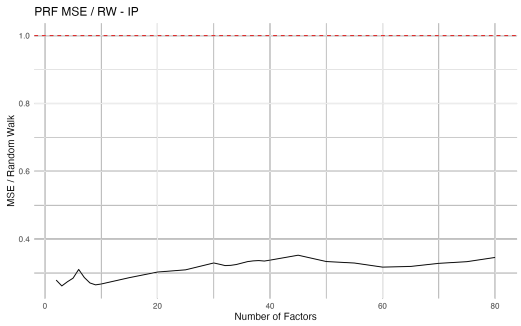
(a) PCA MSFE - CPI



(b) PCA MSFE - IP



(c) 3PRF MSFE - CPI



(d) 3PRF MSFE - IP

Figure 2: MSFE for the PCA and 3PRF and both series

Appendix B. Focus[1]

James Daniel Foltz: Data preparation, Figuring out R commands, The rolling windows mechanism
Francesco Di Marzio: Forcasting with 3PRF
Mino Brescia: Theory of 3PRF
Farzad Yaghoobi: Lasso, Ridge, and PCA theory

1. This refers to the main individual contribution of each group member, but the code and the paper were written jointly since it was unfeasible to treat each method on its own

References

FRED-MD and FRED-QD Databases. https://research.stlouisfed.org/wp/more/2015-012. Accessed on 28/10/2023.

Bryan Kellya and Seth Pruitt. The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, 294:294–316, 2015.

David N Lawley and Adam E Maxwell. Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229, 1962.

Michael W. McCracken and Serena Ng. Fred-md: A monthly database for macroeconomic research. *Journal of Business Economic Statistics*, 2015. URL https://doi.org/10.20955/wp.2015.012. Published as Federal Reserve Bank of St. Louis Working Paper 2015-012.

Christine De Mol, Domenico Giannone, and Lucrezia Reichlinc . Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146:318—-328, 2008.

James H. Stock and Mark W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002. URL http://www.jstor.com/stable/3085839.