Prediction When Factors Are Weak: Supplementary Evidence

Stefano Giglio[*]
Yale School of Management
NBER and CEPR

Dacheng Xiu[†]
Booth School of Business
University of Chicago

Dake Zhang[‡]
Booth School of Business
University of Chicago

This Version: April 6, 2023

Abstract

This paper provides supplementary empirical evidence to "Prediction When Factors Are Weak", Giglio et al. (2022).

[*]Address: 165 Whitney Avenue, New Haven, CT 06520, USA. Email: `stefano.giglio@yale.edu`.
[†]Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. Email: `dacheng.xiu@chicagobooth.edu`.
[‡]Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. Email: `dkzhang@chicagobooth.edu`.

1 Empirical Analysis

In this companion paper we apply the SPCA methodology developed in Giglio et al. (2022) to a standard macroeconomic prediction exercise, using a large set of predictors to forecast inflation, industrial production, and unemployment.

1.1 Empirical Context

Predicting macroeconomic variables like output and inflation is a central exercise in empirical macroeconomics. The availability of large macroeconomic datasets that contain many potentially useful predictors has spurred the application of a variety of methods of dimension reduction to this objective. Some of these methods, like those based on principal component analysis (PCA), reduce the dimensionality of the predictors universe without using information in the target of the forecast (see Stock and Watson (2002)). Others instead use information from the target to help the dimension reduction focus on the most valuable predictors; examples include partial least squares (PLS, Kelly and Pruitt (2015)), targeted PCA (Bai and Ng (2008)), and scaled PCA (Huang et al. (2022)). SPCA belongs to the latter group, as it employs an iterative screening step based on correlation with the target to eliminate useless or noisy predictors.

Because the selection step is designed to eliminate irrelevant predictors (as opposed to downweight them as, for example, PLS does) we expect SPCA to perform best when faced with a large number of predictors that are potentially irrelevant, noisy, or redundant. In our empirical analysis, we therefore explore a context in which a large number of predictors are available to be used for forecasting. Specifically, we include in our set of predictors not only a standard panel of macroeconomic variables, but also a large dataset of individual forecasts of different macroeconomic quantities by professional forecasters. Macroeconomic forecasts have often been included in forecasting exercises, either by using the consensus forecast as an additional predictor (Faust and Wright (2013)) or in the context of optimal forecast combination (Genre et al. (2013)). In our context, we let SPCA decide if and which individual forecasts to use to complement the macroeconomic predictors – so the forecast combination will be decided automatically by SPCA.

1.2 Data

Our empirical exercise combines two datasets. First, we use the standard Fred-Md database (McCracken and Ng, 2016) that contains 127 monthly macroeconomic and financial series.[1] The Fred-Md

[1] The series are grouped in the following categories: output and income; labor market; housing; consumption, orders and inventories; money and credit; interest and exchange rates; prices; stock market. The dataset applies a variety of transformations to the underlying series, which we follow in our analysis. We however make a few adjustments to the series' data transformations, to ensure that all series are stationary and based on economic reasoning. For the Effective Federal Funds Rate (FEDFUNDS), we keep its level (i.e., no transformation) instead of taking the first difference. We also compute the first difference of natural log instead of the second difference of natural log for the following series: M1 Money Stock (M1SL), M2 Money Stock (M2SL), Board of Governors Monetary Base (BOGMBASE; note: starting from the January 2020 (2020-01) vintage, BOGMBASE replaced the St. Louis Adjusted Monetary Base (AMBSL)), Total

data spans the period March 1959 to February 2022. Second, we use individual forecasts from the Blue Chip Financial Forecasts data, which is a monthly survey of experts from various major financial institutions[3] and provides forecasts of interest rates and many other macroeconomic quantities[4] for each of the next six quarters (i.e., current quarter t through $t + 5$), for a total of hundreds of forecasts every month. Our data covers the period February 1993 to February 2022 and we use all forecasts available (for all possible macroeconomic targets) as potential predictors. This gives us up to 18,053 different individual forecasts that could in theory be used as predictors (though, as discussed below, many of these forecasts are available for only a small number of periods, so they are not used in our analysis). Given that the Blue Chip forecast is only available since 1993, we conduct all of our analysis for the period February 1993 to February 2022.

1.3 Out of Sample Forecast Evaluation

We forecast each of the three targets (inflation, industrial production growth, and change in the unemployment rate) using a rolling out of sample procedure. We evaluate the out of sample forecast of SPCA and compare it with two alternative forecasting methods, PCA and PLS. We choose these alternatives because each is a prominent example of a class of methods used in large-dimensional macroeconomic forecasting (respectively, unsupervised and supervised dimension reduction). Each of the three methods we evaluate (SPCA, PCA, PLS) is benchmarked to the forecast of an autoregressive model, whose number of lags is selected by the BIC criterion with a maximum lag of 12 lags, using a direct projection approach (Marcellino et al. (2006), Faust and Wright (2013)). We study forecast horizons of 1 to 12 months.

All of the analysis is performed using a rolling estimation on a 240-months window. At every time t starting at the last month of the window, we predict the cumulated macroeconomic variables from t to $t + h$, where h is the forecast horizon, as in Huang et al. (2022). Within each window, we only keep predictors that have less than 10% missing data points. For those series that are included but

Reserves of Depository Institutions (TOTRESNS), Commercial and Industrial Loans (BUSLOANS), Real Estate Loans at All Commercial Banks (REALLN), Total Nonrevolving Credit (NONREVSL), Finished Goods (WPSFD49207), Finished Consumer Goods (WPSFD49502), Processed Goods for Intermediate Demand (WPSID61), Unprocessed Goods for Intermediate Demand (WPSID62; note: starting from the March 2016 (2016-03) vintage, PPI: Finished Goods (PPIFGS), PPI: Finished Consumer Goods (PPIFCG), PPI: Intermediate Materials (PPIITM), and PPI: Crude Materials (PPICRM) have been replaced with WPSFD49207, WPSFD49502, WPSID61, and WPSID62 respectively), Crude Oil, spliced WTI and Cushing (OILPRICEx), PPI: Metals and Metal Products (PPICMM), Consumer Price Index for All Urban Consumers (CPIAUCSL), CPI: Apparel (CPIAPPSL), CPI: Transportation (CPITRNSL), CPI: Medical Care (CPIMEDSL), CPI: Commodities (CUSR0000SAC), CPI: Durables (CUSR0000SAD), CPI: Services (CUSR0000SAS), CPI: All Items Less Food (CPIULFSL), CPI: All Items Less Shelter (CUSR0000SA0L2)[2], CPI: All Items Less Medical Care (CUSR0000SA0L5), Personal Cons. Exp: Chain Index (PCEPI), Personal Cons. Exp: Durable Goods (DDURRG3M086SBEA), Personal Cons. Exp: Nondurable Goods (DNDGRG3M086SBEA), Personal Cons. Exp: Services (DSERRG3M086SBEA), Avg Hourly Earnings: Goods-Producing (CES0600000008), Avg Hourly Earnings: Construction (CES2000000008), Avg Hourly Earnings: Manufacturing (CES3000000008), Consumer Motor Vehicle Loans Outstanding (DTCOLNVHFNM), Total Consumer Loans and Leases Outstanding (DTCTHFNM) and Securities in Bank Credit at All Commercial Banks (INVEST).

[3]For instance, Bank of America, Goldman Sachs & Co. and J.P. MorganChase.

[4]For instance, the percentage changes in Real GDP, the GDP Chained Price Index, the Consumer Price Index and a set of interest rates (e.g., Federal Funds, 3-month Treasury, Aaa as well as Baa Corporate Bonds).

do have some missing data (mostly Blue Chip forecasts) we forward fill the last non-missing value. About half of the total of around 40 forecasters from BlueChip available in the average month have sufficiently long series of forecasts to be included in our analysis. All predictors are standardized within each window. Then, a forecast is made for $t+1$ using the three different methods, and these forecasts are then joined over time to compute the out-of-sample R^2 (relative to the AR benchmark). When we use the Blue Chip data, we also include dummies for month of the quarter, to account for the fact that the Blue Chip data makes forecasts for calendar quarters irrespective of the month.[5]

Recall that the SPCA procedure presented in Giglio et al. (2022) relies on two tuning parameters, K and $\lfloor qN \rfloor$, whereas PCA and PLS only rely on tuning K. To demonstrate the effect of tuning parameters, we report three versions of the results. We first show the performance of the forecasting methods for different (fixed) number of factors K and different (fixed) choice of $\lfloor qN \rfloor$. In this case, no tuning is needed for SPCA. We then show the performance of SPCA for each K, with a single tuning parameter of SPCA that drives the selection step $\lfloor qN \rfloor$ chosen via 3-fold cross-validation (CV) separately in each time window. Next, we show the results when both the number of factors K (for SPCA, PCA and PLS) and the tuning parameter $\lfloor qN \rfloor$ (for SPCA) are jointly chosen via CV. We consider a range of $\lfloor qN \rfloor$ from 50 to 300.

1.4 Results

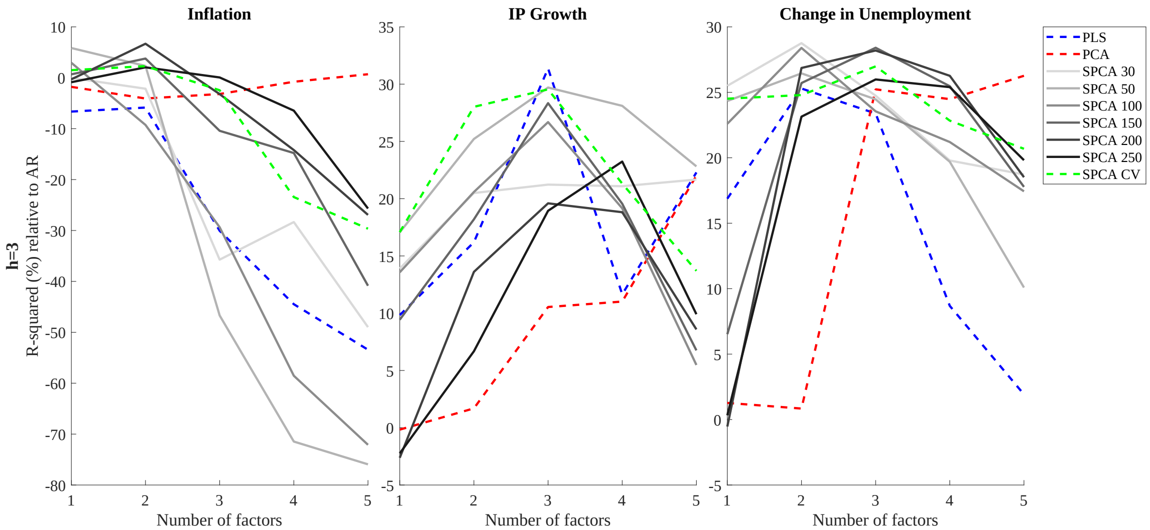
1.4.1 Forecasting Performance

We begin by focusing on prediction at the quarterly (3-month) horizon, which is a standard horizon studied in the literature. Figure 1 reports the out of sample R^2 of different forecasting methodologies relative to the AR benchmark, for inflation (left panel), industrial production growth (center panel), and change in unemployment (right panel). In this figure, the prediction exercise is performed by fixing the number of factors K. For PCA (red line) and PLS (blue line), there are no tuning parameters beyond K. For SPCA, we report separate results for each choice of the tuning parameter K (grey lines), as well as for the value for $\lfloor qN \rfloor$ chosen by CV (green line).

The figure shows several interesting results. First, it is in general hard to predict inflation beyond what an AR model predicts (see also Faust and Wright (2013)): the out of sample R^2s are close to zero or even negative. Only SPCA, among all methods, produces positive R^2s, and it does so using a small number of factors. Predictability beyond the AR model is much higher for IP growth and change in unemployment. Second, the predictive performance of SPCA is generally higher than that of PCA and PLS for most choices of the number of factors. Third, the performance of PCA does depend on the tuning parameter, but in different ways for different targets. For inflation, for example, a lower value of $\lfloor qN \rfloor$ seems to predict better; for industrial production and unemployment, higher values work better. Finally, the performance of all these methods varies quite dramatically with the number of factors, with substantial declines for the methods that use target information

[5] For example, in January, February and March, the "current quarter" forecast always refers to Q1.

Figure 1: OOS Performance of SPCA, PCA and PLS (for different number of factors)



Notes: Each panel reports the out-of-sample R^2 relative to the AR model for a different target, aggregated over 3 months. The three panels predict inflation, industrial production growth and change in unemployment rate, respectively. The green dashed line shows the performance of SPCA with 3-fold cross validation for the tuning parameter $\lfloor qN \rfloor$. The grey lines show the performance of SPCA with fixed number of predictors, $\lfloor qN \rfloor$. The blue dashed line uses PLS. The red dashed line uses PCA. Rolling window of 240 months is used. Sample covers 1993-2022.

(PLS and SPCA with a smaller $\lfloor qN \rfloor$) as the number of factors increases, because of their overfitting issue we explained earlier.

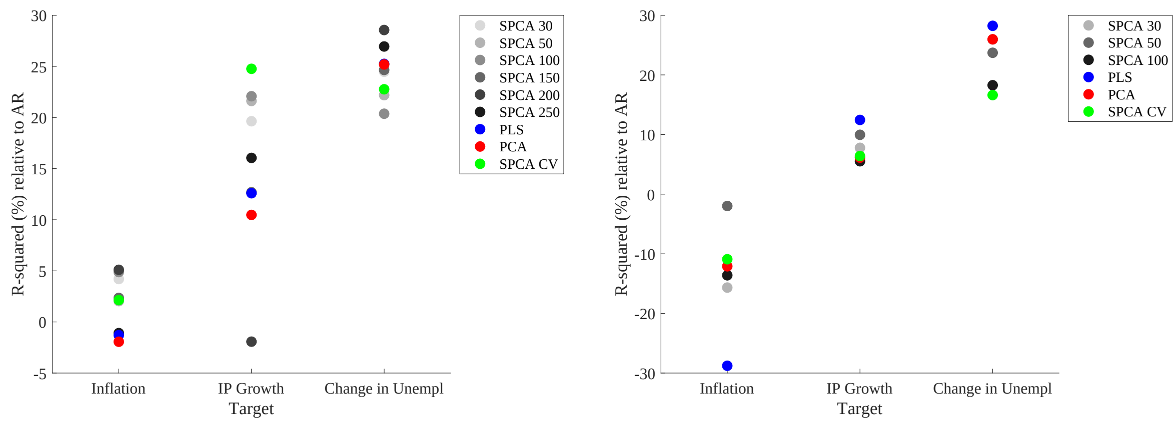
Given how important the number of factors is for the out-of-sample performance, in what follows we choose the number of factors via cross-validation for all three methods (so for SPCA both $\lfloor qN \rfloor$ and K are jointly selected via CV). The left panel of Figure 2 shows the results. Now all three targets (inflation, industrial production growth and change in unemployment rate) appear in the same panel. The panel confirms that SPCA generally performs well in predicting out of sample, doing better than the alternatives (in the case of unemployment, several choices of the tuning parameter $\lfloor qN \rfloor$ outperform PCA and PLS, but not the one chosen by cross-validation). Overall, SPCA tends to do comparatively well when choosing all parameters via cross-validation.

Given the way SPCA chooses the set of predictors, we would expect it to perform best in contexts where there are a large number of predictors, that overall contain valuable information, even if some predictors are redundant or noisy. The forecasting experiment we run here falls in this category: it contains both macroeconomic and financial data (which are likely to contain important individual predictors), as well as a large number of individual forecasts that we would expect to be informative beyond macroeconomic quantities but where a large part of the observed variation is likely dominated by noise. To better gauge the importance of this additional data in the performance of SPCA, the right panel of Figure 2 shows the results of running the same analysis (using the same sample) but

Figure 2: OOS Performance of SPCA, PCA and PLS (using CV to choose the number of factors)

(a) Fred+BlueChip

(b) Fred Only



Notes: The left panel of this figure repeats the analysis of Figure 1, but chooses the number of factors via CV. The right panel performs the same analysis as the left panel, but using only Fred data.

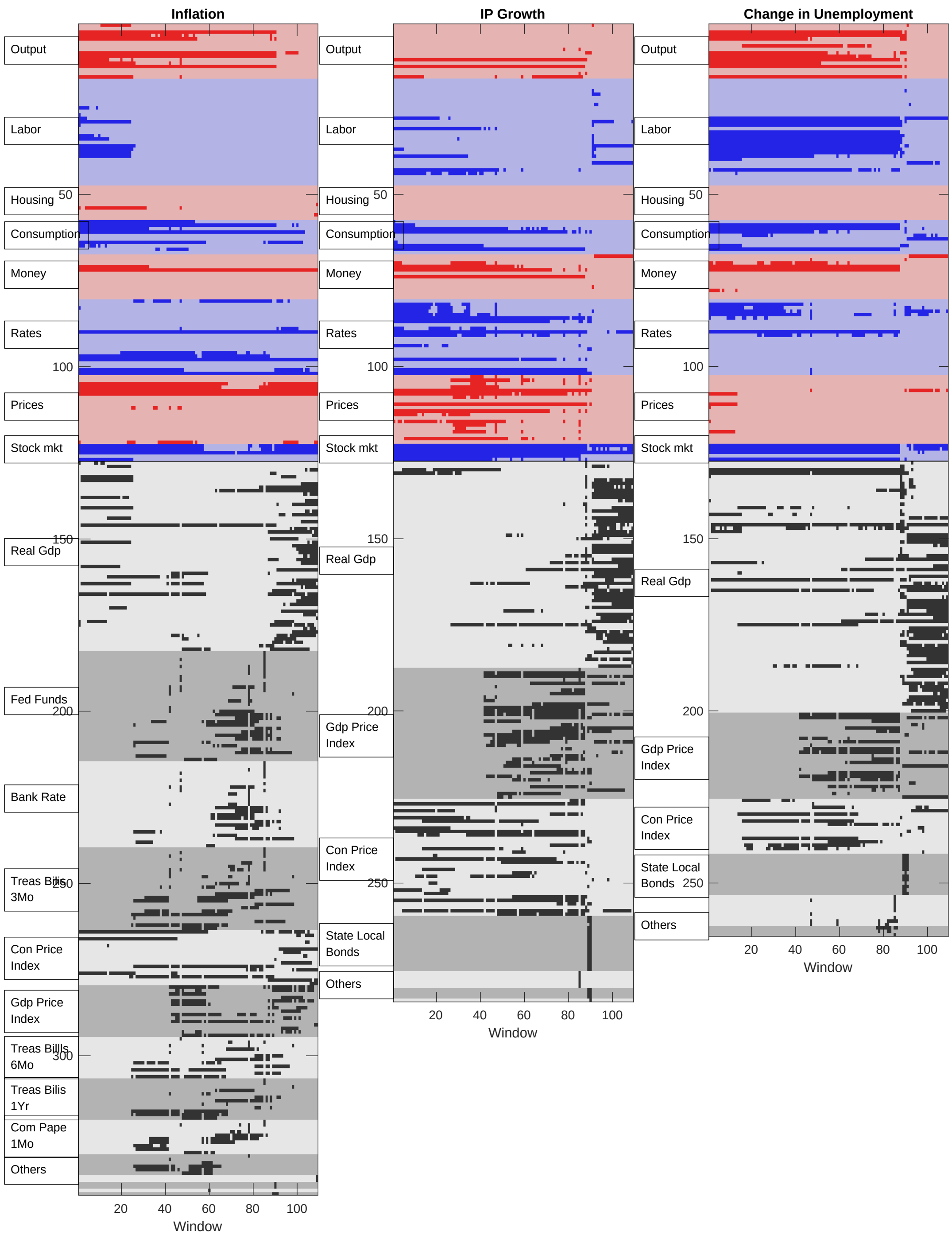
with only the Fred data. The figure shows that while the performance of SPCA remains broadly comparable with the other predictors, it deteriorates compared to PCA and PLS (PLS itself has very mixed performance, though, predicting well IP growth and unemployment, and failing to predict inflation). So, on the one hand, this figure shows that individual expert forecasts are useful for prediction of macroeconomic variables, confirming the results in Faust and Wright (2013); on the other hand, it shows that SPCA does particularly well when working with this large and informative, yet noisy, universe of individual forecasts.

1.4.2 Predictors Selected by SPCA

Next, we study in detail how SPCA selects predictors. Figure 3 shows which variables are chosen by SPCA to extract the first factor (focusing on the 50 with highest correlation with the target, for reasons of readability). For the three targets (one per column), the graph reports which variables were selected in each of the rolling windows in our sample. The top part of the graph collects the 127 Fred variables, grouped according to the standard Fred-Md categorization, in alternating blue and red colors. The bottom part corresponds to the BlueChip surveys, grouped by the target of the individual forecast (therefore, each row in this part of the graph is a forecast of a particular variable, at a particular horizon, by a particular expert). A darker color in this graph means that the variable is selected in that window.

Consider for example the inflation graph on the left. To extract a factor useful to predict inflation, SPCA selects a large number of variables from a few groups: output, consumption, rates, prices, and the stock market. Other groups are almost never selected. Rates are selected more for IP growth, and labor variables are selected more when predicting unemployment. Housing variables are rarely used for all three targets. Note that in many cases, the same predictors from each group are used,

Figure 3: Top 50 Predictors Selected by SPCA



Notes: Under the same settings as Figure 1, each panel visualizes the top 50 predictors selected by SPCA across windows while predicting each target. The first set of variables (in red and blue) are Fred predictors, and the second set (in grey) corresponds to the BlueChip forecasts. For the latter set, only the predictors ever among the top 50 by correlation with the target are visualized.

indicating that the predictive power of these macroeconomic variables is persistent.

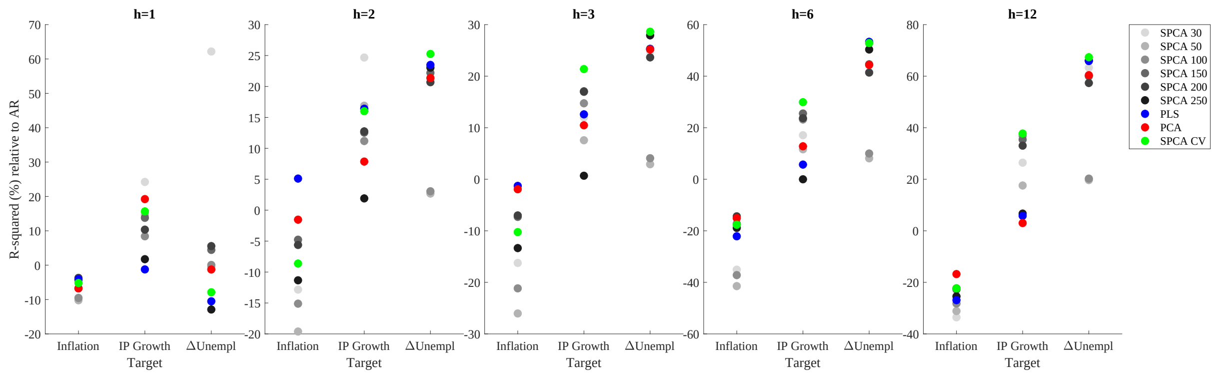
To this macroeconomic set of predictors, SPCA adds a selection of individual forecasts from the BlueChip data as additional predictors. For reasons of space, the greyscale part of the graph shows a subset of these predictors: only those that are selected among the top 50 predictors at least in one window. The graph shows that different types of forecasts are used at different points in time, with some exceptions. Not surprisingly, to predict inflation, forecasts of the consumer price index are always included. To these forecasts, SPCA adds forecasts of GDP in the first and last part of the sample, and interest rates in the intermediate part of the sample. GDP forecasts are used throughout the sample to predict changes in unemployment, and become more dominant for all target variables toward the end of the sample, whereas inflation predictors tend to be more important beforehand. This switch is perhaps due to the fact that in the later part of the sample the zero lower bound was close or binding and inflation was low and not very volatile.

Finally, we note that not all Blue Chip forecasters are the same in terms of forecasting ability. Among the institutions whose forecasts are included in our analysis because they have a sufficiently long time series (each providing tens of forecasts, of different variables at different horizons), we find significant heterogeneity in the frequency with which their forecasts are selected by SPCA. For example, Nomura has its forecasts selected between 23% and 39% of the time at the first iteration (depending on the target). Swiss RE, on the other hand, has its forecasts selected only 0.1% of the time, for each target. This distribution is quite skewed: only 5 institutions have their forecasts selected more than 10% of the time for each target, out of the 20 included in our sample. Similar results hold when looking at selection at any iteration of SPCA.

1.4.3 Joint Forecasts using Many Targets

Next, one special feature of SPCA is that it can operate the selection using a set of multiple targets jointly. In fact, using multiple targets is required by the theory (see Giglio et al. (2022)) to do inference, as long as there are more than one factors in the true DGP. We implement this here by predicting each target at horizons of 1, 2, 3, 6 and 12 months jointly. Figure 4 reports the out of sample R^2s on each horizon. There are two main results that this figure highlights. First, SPCA tends to do on average well at longer horizons (3, 6 and 12 months), whereas its performance is more uneven at shorter horizons. Second, comparing the middle panel (predicting one quarter ahead) with the left panel of Figure 2, which focused on the 3-month horizon only, we see that the use of other horizons to help select predictors has different effects for different targets. It significantly improves the forecasting ability for unemployment, but reduces the forecasting ability for IP growth (mildly) and inflation (significantly so). Overall, the performance of SPCA remains on par with the other predictors when using multiple targets, especially at longer horizons.

Figure 4: OOS Performances - Different Targeted Horizons



Notes: Similar to Figure 2, but showing the out of sample R^2s at different horizons, and using all the horizons concurrently to estimate the factors in SPCA.

1.4.4 Time Series of the Forecasts

Finally, we study the time series of our out-of-sample forecasts at different horizons, using the estimates obtained in Section 1.4.3, for horizons of 1, 2, 3, 6 and 12 months. Figure 5 reports SPCA's forecasts with asymptotic forecast standard errors at each maturity. In the figure, the blue dots represent the underlying time series that is the target of the forecast: log CPI, log IP, and unemployment, all scaled to start from 0 at the beginning of the sample. For readability, we show the forecasts every six months, each for horizons up to 12 months. Standard errors are obtained using the asymptotic distributions derived in Giglio et al. (2022), and are plotted in three shades (the 10th and 90th percentiles in the darkest shade, 5th and 95th in the middle shade, and 1st and 99th in the lightest shade).

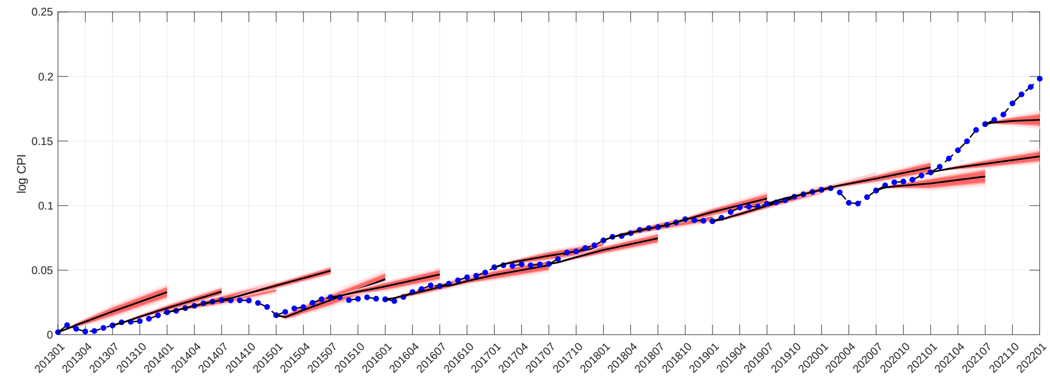
Overall, SPCA does a good job forecasting the three series, with the forecasts often anticipating changes in the direction of the different variables. For example, IP forecasts predicted the increase starting in 2016, and the decrease that started in 2018. Of course, in other times the forecasts miss significantly, sometimes for several periods in the same direction. Two examples: first, forecasts do not fully anticipate the persistent decrease in unemployment that occurred during 2013 and 2014. Second, all forecasts miss (as they should have) the unexpected and extraordinary events of the Covid pandemic (both the initial shock and the recovery). In that period, the point estimates change dramatically over a short period of time, and standard errors increase noticeably, demonstrating the large amount of uncertainty about the path of the economy during those times.

References

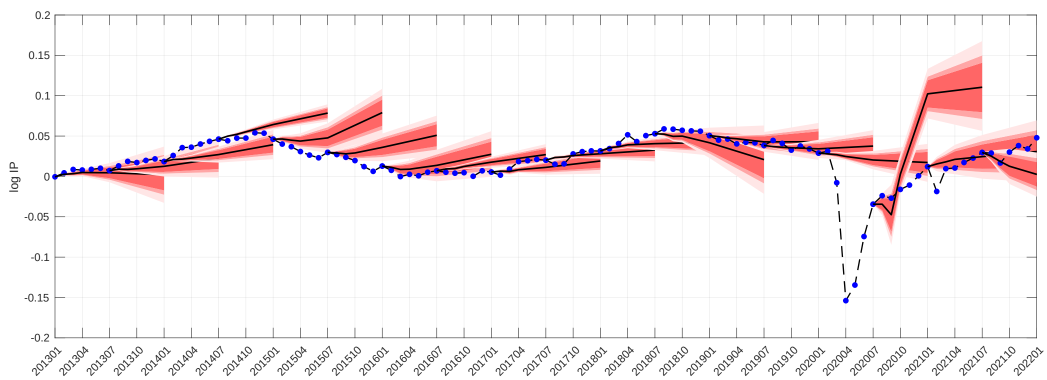
Bai, J. and S. Ng (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics 146*(2), 304–317.

Figure 5: Fan Charts

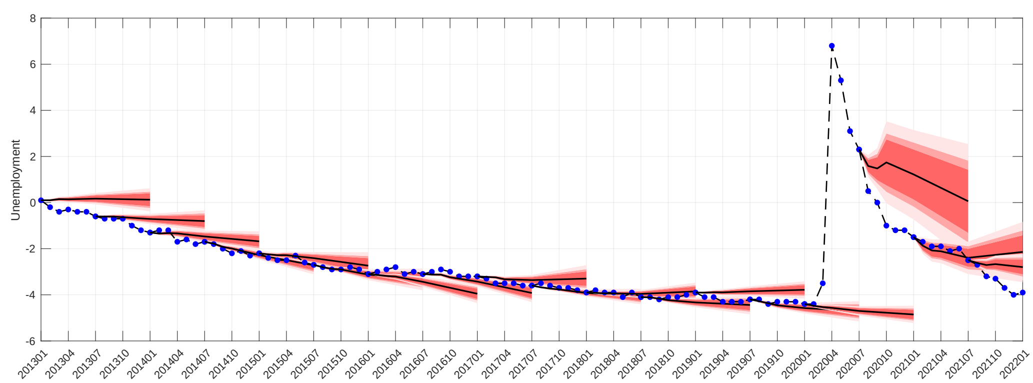
(a) Inflation



(b) IP Growth



(c) Change in Unemployment



Notes: Using the same estimates as Figure 4, each panel shows the forecasts and confidence intervals for horizons up to 12 months. The forecasts are shown every 6 months, in alternating red and green colors (for readability). The blue dots are the cumulative targets of the forecasts.

Faust, J. and J. H. Wright (2013). Forecasting inflation. In *Handbook of economic forecasting*, Volume 2, pp. 2–56. Elsevier.

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

Giglio, S., D. Xiu, and D. Zhang (2022). Prediction when factors are weak. Technical report, University of Chicago.

Huang, D., F. Jiang, K. Li, G. Tong, and G. Zhou (2022). Scaled pca: A new approach to dimension reduction. *Management Science 68*(3), 1678–1695.

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

Marcellino, M., J. H. Stock, and M. W. Watson (2006). A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of econometrics 135*(1-2), 499–526.

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

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